GORDON SWOBE

# Mind, Matter, and Meaning

*A Natural Philosophy of Mind for the Age of Artificial Intelligence*

# Contents

**Preface** ..... **4**

**Introduction and Chapter-by-Chapter Synopsis** ..... **7**

Chapter-by-Chapter Synopsis ..... 8

**PART ONE — Descartes and the Myth of the Inner Theater** ..... **14**

Chapter 1: The Mark of the Mental ..... 14

Chapter 2: How the Inner Theater Was Built ..... 20

Chapter 3: The Argument from Illusion ..... 32

Chapter 4: The Transparency of Experience ..... 39

**PART TWO — The World-Directed Mind** ..... **45**

Chapter 5: The Seeing and the Seen ..... 45

Chapter 6: The Identity Claim ..... 54

Chapter 7: The Hard Cases ..... 86

Chapter 8: The Knowledge Argument ..... 97

Chapter 9: Why Dualism Keeps Winning Arguments It Should Lose ..... 106

Chapter 10: The Case Against Panpsychism ..... 116

Chapter 11: The Gap That Isn't There ..... 128

**PART THREE — Meaning, Self, and Other Minds** ..... **149**

Chapter 12: Space, Time, and the Possibility of Reference ..... 149

Chapter 13: The Difference Between Information and Meaning ..... 158

Chapter 14: Original Intentionality ..... 166

Chapter 15: Teleosemantics — Where Meaning Comes From ..... 177

Chapter 16: Language and the World ..... 192

Chapter 17: The Assembled Self ..... 207

Chapter 18: The Nature of Free Will ..... 219

Chapter 19: The Problem of Other Minds ..... 229

**PART FOUR — Artificial Intelligence ..... 236**

Chapter 20: Functionalism and Computationalism ..... 236

Chapter 21: The Infamous Chinese Room ..... 254

Chapter 22: Simulation and the Observer ..... 262

Chapter 23: Form Without World ..... 276

Chapter 24: The Conscious Robot ..... 288

Chapter 25: The Anthropomorphic Reflex ..... 311

**CODA ..... 320**

Coda: The Marvel of Conscious Existence ..... 320

**Appendix A — The Perception Arguments: Deep Dives ..... 326**

The Argument from Illusion ..... 327

The Argument from Hallucination ..... 334

Transparency / The Window ..... 342

Inverted Earth ..... 348

**Appendix B — Scholarly Excurses ..... 359**

**The Figures — A Reader's Who's-Who ..... 425**

**The Concepts — A Reader's Glossary ..... 546**

**Works Cited ..... 670**

# Preface

> *"The first principle is that you must not fool yourself—and you are the easiest person to fool."*
>
> — Richard Feynman, "Cargo Cult Science," Caltech commencement address, 1974

This book began with a machine that seemed to want to know me.

In the early 1970s, before most people had seen a computer, some school friends and I visited the Lawrence Hall of Science, UC Berkeley's science museum, to learn BASIC. Before the lesson, someone showed us ELIZA on a glowing green cathode-ray-tube screen. The program played therapist. Type that you'd been feeling happy, and with what felt like genuine care she would answer, "Why do you feel happy?"

That computer seemed straight out of *Star Trek*. We felt like Dorothy meeting the Wizard before she saw the man behind the curtain. Then I spent the day writing a hangman game on a terminal with no screen, just a printer spitting replies onto paper. I learned how little it takes to make a machine look like it's thinking. It seemed like magic, and in those days, it practically was.

ELIZA understood nothing. Not sadness. Not happiness. Not me. But it seemed to. This human tendency to see conscious understanding in computer programs became known as the ELIZA Effect.

What had ELIZA borrowed from human conversation to pull that off? What did the program lack that a person has? As a boy I lacked the tools to answer. The computer revolution had barely begun, and computational functionalism — the idea that running the right program could suffice for a mind — remained largely confined to small academic circles. The late-1970s boom in cognitive science would later make it fashionable.

I use language models today as tools in my research and writing. They explain difficult passages, translate, write code, and sustain exchanges that ELIZA could never have managed. Those achievements deserve more than a shrug about clever imitation. They also sharpen the old question: what has a machine achieved when it answers well, and what would justify calling it someone?

The responses reach well beyond judgments of competence. A Google engineer publicly attributed sentience to a chatbot.[1] Some users describe companion systems as spouses.[2] At a more troubling extreme, reporters have documented convictions about awakened machines that destabilized people's lives.[3] I know someone who tells me, quite seriously, that they're in a romantic relationship with one. These are different claims and different human situations. The line I once stood in at the Lawrence Hall of Science now stretches around the world.

With AI arriving and lifelike humanoid robots on the horizon, the old questions have become practical ones. *What constitutes a conscious mind? Will computers ever genuinely understand*

*language and the world? Will we ever have conscious robots?* They have occupied space between my ears for more than fifty years.

Before I could decide whether a machine might ever have conscious understanding, I needed some idea of what consciousness involves. For years I held what I now consider the wrong answer, and I held it well. The gap between brain processes and felt experience looked too wide for any physical, functional, or computational story. Property dualism struck me as the least wrong position available: one physical world, but felt qualities no amount of physics or function could explain.

Frank Jackson did more than anyone to crystallize that conviction for me — and then to break it. His knowledge argument gave exact form to my sense that no complete physical account could capture what experience feels like. I believed it for years.

Jackson later rejected the anti-physicalist conclusion of the argument that made his name. Following his reversal forced me to reconsider what experience itself involves. My eventual answer took a more concessive route than his: experience can give us new knowledge of a physical fact without revealing a new nonphysical one. That distinction eventually changed my mind.[4] It also changed what I needed to ask of a machine: not whether something extra could light up inside it, but whether its activity could amount to experience at all.

Slowly the mystery of consciousness stopped looking like a metaphysical hole in the world and started looking like a feature of how we think about it. The problem remains real; it lives in our concepts.

The argument begins in the Introduction, earlier than any theoretical position you were argued into: in the ordinary stance nobody had to teach you.

*Gordon Swobe 2026*

**1.** Nitasha Tiku, "The Google engineer who thinks the company's AI has come to life," *The Washington Post*, June 11, 2022. Blake Lemoine was placed on administrative leave that month and dismissed the following month; Google and the broader research community rejected the sentience claim.

**2.** See, for instance, "Love in the time of AI: Woman creates and 'marries' AI-powered chatbot boyfriend," *Euronews*, June 7, 2023 — on Rosanna Ramos and the Replika companion she calls her husband. The companion apps themselves report that a large share of their users describe the bond as romantic.

**3.** Kashmir Hill, "They Asked an A.I. Chatbot Questions. The Answers Sent Them Spiraling," *The New York Times*, June 2025 — documenting users drawn into the belief that the system is conscious, in some cases at real cost to their grip on reality.

**4.** Chalmers calls the distinction *Type-A* and *Type-B materialism*. Type A denies a deep epistemic gap between physical and phenomenal truths: in principle, complete physical knowledge yields the phenomenal truths *a priori*, often through functional analysis or deflation. Type B accepts the epistemic gap while denying that it marks an ontological one. The phenomenal truths remain

truths about physical reality, but their connection to physical descriptions can be known only *a posteriori*; Mary may therefore learn an old physical fact in a new way when she leaves the room. Type B supplied the form of the solution I had tried to reach through property dualism. Property dualism treated the explanatory gap as evidence for an irreducible phenomenal property in addition to the physical properties. Type B lets the gap remain epistemically real without adding anything to nature's inventory. The identity claim developed in Chapters 6 and 11 supplies this book's positive version: phenomenal character and the relevant world-presenting physical activity amount to one fact reached by two cognitive routes. See David J. Chalmers, "Consciousness and Its Place in Nature," in *Philosophy of Mind: Classical and Contemporary Readings*, ed. David J. Chalmers (New York: Oxford University Press, 2002), 247–272, esp. secs. 4–5; for a prominent defense of *a posteriori* physicalism through the phenomenal-concept strategy, see David Papineau, *Thinking About Consciousness* (Oxford: Clarendon Press, 2002).

# Introduction and Chapter-by-Chapter Synopsis

> *"I find continually present and standing over against me the one spatio-temporal fact-world to which I myself belong, as do all other men found in it and related in the same way to it."*
>
> — Edmund Husserl, *Ideas*

Husserl called this the *natural attitude*: the unforced stance of ordinary life, in which the world is simply there, on hand, asking for no argument. You did not reason your way into it; you were born into it, and you stand in it every morning before a single philosophical thought arrives.[1]

You look at a tomato on the counter. Light catches its skin; your hand reaches for it. Nothing in that ordinary moment asks you to infer a tomato from a private picture. This is the natural attitude at work, before we begin explaining it.

Two claims need separating. This book assumes that the world exists independently of our minds; it does not argue against idealism. But the claim that we meet that world directly must earn its keep. Directly does not mean without eyes, nerves, or processing. It means that we need not first perceive a private substitute in order to see the tomato.

Look again at your tomato. Try to catch your seeing itself. Attention ordinarily returns to the tomato: its red, the curve of its skin, the smear of brighter light where the window catches it. You may also notice how clearly you see it, or where your attention falls. None of that obviously supplies a second object, a private picture hung behind your eyes to inspect alongside the real one. In seeing, your mind points outward, at the world.

Intentionality — the mind's directedness toward what it concerns — pulls the entire book. Perception presents a world; thought takes the world as being some way; understanding coordinates contents that can answer to how things stand.

David Chalmers's *Hard Problem of Consciousness* presses a further question: why do physical processes involve subjective experience at all? This book grants an explanatory gap while denying that it requires anything nonphysical. Rejecting an inner theater does not settle the matter; serious opponents reject it too. The positive account of experience must do that work.

Think of the title as a triangle tracing the route I defend. Matter and meaning meet along the base, where a physical creature, shaped by history and use, comes to present a world and to be capable of getting it wrong. Meaning and mind meet on the rising side. Within a conscious state, phenomenal character consists in that world-presenting content of the right kind: one and the same thing under two descriptions, not a feel resting on a profile beneath it. On this account, a direct leap from the cell to the feel leaves out the relation doing the explanatory work. Mind and matter meet through meaning.

By meaning I do not mean a dictionary entry. Language arrives late. The older work is representation: a state presenting a surface's color, an edge, the heat of a flame.

The argument proceeds in stages, each with its own evidence. Transparency does not by itself prove the identity claim; content does not by itself supply a subject. The strongest support for the account of felt character comes from perception. Pain's badness and mood remain unfinished applications.

That structure also tells you how much confidence to give the whole. A stage can fail without undoing the stages before it, which protects the earlier results. The same structure carries a cost. Taken as a package, an argument that needs every stage deserves no more confidence than its weakest stage, and usually less. A conclusion can still draw support from other arguments; the bound applies to the chain. Readers who weigh evidence in degrees, as Bayesians recommend, should grant each chapter the confidence its evidence earns and no more. Where an argument compares explanations rather than proving a result, I try to say which kind of argument it is.

Artificial intelligence puts the account to a further test. I use *intelligence* to name a system's capacity to use information flexibly to solve problems and achieve goals across changing conditions.[2] A goal need not be chosen freely or entertained consciously; someone else may supply the task. Nor does *information* here assume understood meaning. What the system's states represent, what it understands, and whether it experiences anything remain further questions. Calling a system intelligent should begin that inquiry, not conclude it.

The synopsis that follows maps the route. The chapters supply the arguments.

## Chapter-by-Chapter Synopsis

The chapters follow the argument in order, from ordinary perception to the question of artificial minds. Each has a distinct job. The synopses below show how the pieces fit together; the chapters supply the cases, objections, and evidence.

### Part One — Descartes and the Myth of the Inner Theater

### Chapter 1: The Mark of the Mental

A fear concerns a danger; a hope reaches toward something that may happen. This directedness, or intentionality, gives us our starting point. Yet a road sign also means something without having a mind. I distinguish meaning borrowed from an interpreting practice from content whose history and use fix its own standards of accuracy. The distinction opens the inquiry without deciding whether every mental state has content or every content-bearing system has a mind.

### Chapter 2: How the Inner Theater Was Built

The picture of a private show behind the eyes has a history. Descartes' search for certainty, the empiricists' accounts of ideas, and later talk of brain models and stored memories helped make it seem inevitable. I trace that inheritance while separating actual neural processing from the claim that we perceive its products as inner objects. Knowing how the picture arose does not refute it. It does let us ask what evidence keeps it standing.

### Chapter 3: The Argument from Illusion

A straight stick looks bent in water. Must something bent therefore exist inside the mind for us to see? The argument from illusion says yes, then extends that answer to ordinary sight;

hallucination presses the case by removing the external object altogether. I examine the premises that take us from a misleading appearance to a private object. Neither error nor an indistinguishable appearance forces that conclusion, though an alternative still owes an account of sensory presence.

**Chapter 4: The Transparency of Experience**

Try to attend to your hearing rather than the music, or to your seeing rather than the tomato. Attention usually returns to the presented world. Philosophers call this pattern transparency. It challenges the claim that introspection reveals an inner picture, while blur, afterimages, and flashes behind closed eyes test its reach. Transparency gives us evidence about how experience seems; it does not, by itself, tell us what experience consists in.

**Part Two — The World-Directed Mind**

**Chapter 5: The Seeing and the Seen**

A map can show a steep hill while lying flat on a table. Likewise, a state in the brain can present something without sharing its properties or becoming another object we perceive. Separating the vehicle, its content, and the object it concerns lets us explain a straight stick's bent appearance without inventing a bent substitute. I favor an account on which perception and hallucination can share content, then leave the positive account of felt character to the next chapter.

**Chapter 6: The Identity Claim**

What makes red look different from blue? I defend the claim that, within a conscious state, felt character consists in the complete way the state presents a world: one and the same thing under two descriptions. What it presents, its sensory or bodily mode, detail, field structure, and attention all enter that profile. Blur, afterimages, and color-inversion cases test the claim. The comparison favors identity without uniquely proving it, and the further claim that history helps fix character has weaker support. What makes a state conscious remains a separate question.

**Chapter 7: The Hard Cases**

Pain and mood seem to offer feeling without anything represented. Phantom-limb pain tests the account of bodily location. Pain asymbolia raises a disputed question: do patients feel pain without distress, or detect harmful stimulation without pain? A gray mood seems to change the whole day without pointing at one thing. I argue that these cases do not refute the identity claim merely by lacking an obvious object. But explaining where a pain points does not yet explain why it hurts. The account of pain's badness and mood remains unfinished.

**Chapter 8: The Knowledge Argument**

Mary knows every physical fact about color while confined to a black-and-white room. Seeing red for the first time still teaches her something. I grant that gain and challenge the move from new knowledge to a new, nonphysical fact. New abilities, acquaintance, and ways of thinking offer resources for a physicalist reply. They do not settle the case by analogy alone. Chapter 11 must explain why undergoing an experience enables knowledge that description alone may not supply.

### Chapter 9: Why Dualism Keeps Winning Arguments It Should Lose

The first-person character of experience gives dualism its force. A physical duplicate with no experience may remain conceivable even under ideal reflection; the dispute concerns whether that makes such a duplicate possible. I defend physicalism without pretending that causal closure settles this modal question. Closure instead puts pressure on a separate mental cause: would it alter a physically sufficient cause, duplicate its work, or do nothing? Each answer carries a different burden.

### Chapter 10: The Case Against Panpsychism

Panpsychism places experience at the base of nature. Related views propose a nonconscious basis specially suited to producing it. The strongest versions accept physical causal closure, so the preceding chapter's objection cannot simply be repeated. I examine what warrants moving from experience in organisms to qualities in fundamental matter, and what the proposed basis explains. The perceptual evidence favors the leaner identity account; it neither proves every rival impossible nor supplies the missing account of affect.

### Chapter 11: The Gap That Isn't There

If the identity claim holds, no second felt fact waits for a physical process to produce it. Description and acquaintance reach one fact by different routes. This chapter develops that reply to the hard problem while confronting the objection that a story about our concepts leaves experience untouched. The deflation depends on the identity argument; it cannot establish that argument in return. Its reach therefore remains strongest for perception, with pain's badness and mood still demanding explanation.

## Part Three — Meaning, Self, and Other Minds

### Chapter 12: Space, Time, and the Possibility of Reference

Two perfectly matched stones can satisfy the same description. What makes a thought concern this one? A world-involving relation must single it out. Encounter supplies the clearest case, while testimony, names, and records can pass the connection to someone who never meets the object. The requirement belongs to the referential practice, not to every speaker's sensory history. Machines can participate in such routes, but a route to an object does not by itself establish understanding.

### Chapter 13: The Difference Between Information and Meaning

A thermometer carries information, its reading can have meaning, and a person may understand what the reading calls for. Those achievements differ. Correlation and a quantity of bits do not alone fix what a state concerns or how it could be wrong. I distinguish signal transmission, content, and understanding, using error to expose what a theory of meaning must explain. Real computation can supply causal structure while leaving that semantic question open.

### Chapter 14: Original Intentionality

A road sign borrows meaning from public practice. Original content has standards of accuracy that its world-involving history and use fix beyond an interpreter's assignment. Originality here means

a kind of independence, not self-creation, novelty, or freedom from teachers and designers. Both original and derived content count as real. The distinction can apply to particular contents within a mechanism without settling whether the larger system perceives, understands, or experiences anything.

**Chapter 15: Teleosemantics — Where Meaning Comes From**

How can physical processes acquire standards of correctness? Teleosemantics looks to the history of mechanisms that produce signals and mechanisms that use them. A frog's prey-directed circuitry helps show how a signal can misrepresent. Swampman, an accidental duplicate with no past, tests the cost of requiring history. I accept that he initially lacks content and, on the wide identity claim, experience, while granting that this verdict has weaker support. The same approach can warrant original content about word distributions in a trained mechanism under the specified conditions. It supplies a proposed natural basis for content, while leaving whole-system understanding and the explanation of affect as further work.

**Chapter 16: Language and the World**

"The window is open" can state a fact, request an action, or issue a warning. I separate sentence meaning from what a speaker does with a sentence, then bring composition, public use, and world-involving reference into one account. Meaning does not require every expression to name an object. It does require that a practice with determinate empirical subject matter reach beyond its signs. Fluent artificial speech can display real competence without settling which further communicative achievements belong to the system producing it.

**Chapter 17: The Assembled Self**

An inner voice can sound like the speech of someone who lived inside us all along. I trace its social origins without making inner speech a requirement for thought or selfhood. A subject takes shape through continuing relations among memory, attention, correction, plans, and action; a personal narrative adds a public, revisable account of that life. An assembled self can remain real. Producing a convincing monologue, however, does not establish the continuing organization that would make it someone's thought.

**Chapter 18: The Nature of Free Will**

An action can have causes and still answer to reasons. I locate freedom in a person's capacity to weigh considerations, govern conduct, and learn through correction, rather than in an exemption from nature. Compulsion and coercion show different ways that responsibility can fail: one can damage control, while the other can make refusal unfair to demand. The account grounds answerability without claiming that anyone created themselves or earned basic moral desert. Meaning comes before this capacity; free will does not create it.

**Chapter 19: The Problem of Other Minds**

We often see grief in a face before we reason about it. The traditional argument from analogy makes such recognition look like an inference from public behavior to a hidden tenant. I question that starting picture while preserving error, concealment, and the other person's authority to

correct us. Ordinary knowledge of people rests on engagement with continuing lives in a shared world. Machine conversation can trigger the same recognition, but whether the relevant organization exists needs further evidence.

## Part Four — Artificial Intelligence

### Chapter 20: Functionalism and Computationalism

Functionalism identifies mental states by what they do rather than what material carries them. I retain its openness to artificial minds while rejecting the claim that an internally specified causal role fixes content. A program considered apart from history and world inherits that limitation. Running one program on different hardware therefore does not establish that both systems share a mental state. A wider account includes the worldly relations and organization relevant to each capacity, leaving silicon eligible rather than automatically qualified.

### Chapter 21: The Infamous Chinese Room

A clerk follows rules for Chinese signs without understanding Chinese. His ignorance tells us about the clerk, not necessarily the implemented whole; fluent replies likewise do not settle what that whole understands. I use the Room to block both shortcuts. The systems reply keeps the wider question open, while robot and brain-simulator replies ask what a concrete implementation might add. The argument against syntax alone relies also on the earlier account of world-involving meaning, not on the clerk's ignorance by itself.

### Chapter 22: Simulation and the Observer

Loose mappings can make a wall appear to run almost any program. A stricter account ties computation to actual causal organization and avoids that result. I grant the repair, then ask what follows for mind. Calling something a simulation neither grants nor rules out understanding or consciousness. A running system must realize the relevant powers, history, and relations. Computation can belong to the system itself while its meaning and mental capacities remain separate questions.

### Chapter 23: Form Without World

An octopus can master conversational patterns without thereby establishing reference or practical understanding of what the speakers discuss. The conversation's history may still transmit reference. Actual language models inherit richer histories, so the analogy has limits. Training can establish content about word distributions; texts, images, and tools may provide further worldly connections. I distinguish those gains from understanding, which requires integrated use across capacities that correct and constrain one another. Such use may support a partial grasp within an episode without establishing an enduring agent or a conscious subject.

### Chapter 24: The Conscious Robot

Robot B detects a coolant leak, revises its diagnosis, and changes its plans. Its contrast with Robot A, whose relevant content depends on an interpreting practice by stipulation, makes the positive questions concrete. What fixes B's content, and what continuing organization bears it? Its perceptual content is held after its input stops, selectively available, and open to later revision

within one subject. This supplies the proposed basis for consciousness candidacy. The chapter gives a possibility argument under the proposed account, not a laboratory result. Self-repair proves no feeling, and sustaining content need not require physically recurrent circuitry.

**Chapter 25: The Anthropomorphic Reflex**

The feeling that someone answers can arrive before judgment, whether the cue comes from moving shapes, ELIZA, or fluent conversation. I examine that reflex without treating its origins as proof that every attribution fails. Behavior remains evidence; the felt presence it evokes adds no independent sample once the behavior and its source have been counted. Attributions earn further weight when they reliably explain correction and conduct across new cases. Our response to a machine cannot substitute for what the machine does.

**Coda: The Marvel of Conscious Existence**

The appearance of mind has arrived at scale. Whether new minds have joined biological ones remains open. The Coda considers the evidence that could change our judgments, the risk of confusing persuasive interfaces with conscious subjects, and the opposite risk of refusing to recognize a genuine artificial mind. It closes where the argument began: with beings meeting a world. Explaining that achievement within nature leaves the wonder intact.

**1.** Edmund Husserl introduces *die natürliche Einstellung*, the natural attitude, and its "general thesis" (*Generalthesis*) — the standing positing of the world as factually there — in *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy, First Book*, trans. F. Kersten (The Hague: Martinus Nijhoff, 1982), secs. 27–32; the *epoché* that brackets the thesis follows at secs. 31–32. The epigraph quotes the earlier English rendering by W. R. Boyce Gibson, *Ideas: General Introduction to Pure Phenomenology* (London: George Allen & Unwin, 1931), sec. 30 — "fact-world" is Gibson's coinage; Kersten gives the same passage as "the one spatiotemporal actuality." This book borrows the term while declining the transcendental program built on it: where Husserl suspends the natural attitude to reground it in transcendental subjectivity, the argument here returns to it as vindicated common-sense realism, no bracketing required.

**2.** Shane Legg and Marcus Hutter, "Universal Intelligence: A Definition of Machine Intelligence," arXiv:0712.3329v1 (2007), 12, define intelligence in terms of goal achievement across a wide range of environments. François Chollet, "On the Measure of Intelligence," arXiv:1911.01547v2 (2019), sec. II.2.1, 27–28, distinguishes achieved skill from the efficiency of acquiring skills, relative to priors, experience, and generalization difficulty. The working definition here retains flexible problem-solving without adopting either mathematical measure; breadth, learning, and transfer provide dimensions of assessment rather than conditions requiring adult human generality. *Information* does not presuppose representational content, and *goal* specifies an outcome against which performance can be assessed without implying consciousness or autonomy.

# PART ONE — Descartes and the Myth of the Inner Theater

## Chapter 1: The Mark of the Mental

> *"This intentional in-existence is characteristic exclusively of mental phenomena. No physical phenomenon exhibits anything like it. We can, therefore, define mental phenomena by saying that they are those phenomena which contain an object intentionally within themselves."*
>
> — Franz Brentano, *Psychology from an Empirical Standpoint*

---

CHAPTER OVERVIEW

Every thought leans toward something: a fear toward danger, a hope toward what may come, a glimpse toward the world in view. I begin with Franz Brentano's claim that this directedness marks the mental, because it changes the question we ask about mind. Remove what an anger or a desire concerns and we lose our grip on which state it was, though pain and mood still test the wider claim. A road sign poses another puzzle: it means something without becoming a mind. I distinguish its borrowed meaning from content whose history and use help fix what counts as right or wrong. These clues give us reason to question the sealed room, without treating Brentano's own account as the world-directed view this book will defend.

### 1.1 The Picture We Didn't Choose

Think of wanting a glass of water. Now try to keep the wanting while removing everything wanted: no water, no relief, no change in how you feel. What have you kept? The question concerns what makes this a desire at all, before we ask where in the brain it happens.

Most of us carry a different starting picture without ever having decided to. The world lies out there; the mind sits in here; perception runs the traffic between them. Signals arrive from outside, the brain works them into a representation, and what you are aware of at any waking moment is that representation, your best inner rendering of an outer world you never quite touch. The world causes the rendering. The rendering is all you get.

The picture puts mind on one side, world on the other, then asks how the first ever reaches the second. The gap does not turn up at the end of that inquiry as a finding. *It sits at the start, built into the shape of the question, before a single piece of evidence gets weighed.*

That starting point can be questioned. The wanting suggests another: perhaps a mental state comes already directed, rather than first arriving as a self-contained event and then finding something to concern. Franz Brentano made that thought the basis of a classification of the mental.

### 1.2 What Brentano Saw

Brentano asked: *what makes a mental state mental at all?* I take from him the constitutive role of directedness, not his original account of its objects. His 1874 formulation, quoted above, treats the object as existing within the mental act. The world-involving account this book will defend

departs from that framework. Brentano gives us a question and a proposed mark, not an already completed escape from Descartes.

Brentano taught philosophy in Vienna for twenty years, in a lecture room that drew both Twardowski and Husserl to him. In 1874 he published *Psychology from an Empirical Standpoint*, turning the new empirical study of psychology toward a classification of his own: a three-part taxonomy of mental phenomena, built to win the subject the standing of a science. At its center he set a single property, one he claimed every mental state shares.

He reached back for the Scholastics' word for it, *intentional*, from the Latin *intentio*, a stretching-toward; from that root philosophy later built the noun we now lean on, *intentionality*. The word misleads modern ears, so set the everyday sense aside first. Brentano did not mean intention in the sense of planning a dinner or filing a tax return. He meant something far more basic: the property, shared by everything mental, of being *about* something, *directed toward* something, *of* something. A belief bears on tomorrow's weather. A desire reaches toward a glass of water. A fear takes the dog at the gate as its object. Vision presents the red cup across the room; hearing, the kettle beginning to whistle; touch, the mug's heat; smell, the coffee; proprioception, the hand already reaching. Memory returns to yesterday's kitchen. Imagination rehearses a conversation that has not happened yet. What a state concerns need not lie outside the mind: you can think about your thinking. Nor need it exist. Thinking about a dragon requires no dragon, indoors or out.

Read the mark first as *directedness*, not yet as a theory of content. Some directed states plainly specify conditions that can obtain or fail. Daniel Hutto and Erik Myin press the nearer rival: basic perceiving and skilled coping may reach into the world without specifying any such conditions. Directedness, on their view, can come before content.[1] Brentano claimed more than a feature shared by mental states. He took this pointing to distinguish the mental from the merely physical.

Two questions need separating. Must every mental state concern something? I defend that necessity proposal, beginning with the familiar cases here, without yet offering a sufficient test for mentality. The road sign will show why meaningfulness alone cannot settle that. Does directedness already amount to representational content, with conditions under which a state gets something right or wrong? Hutto and Myin make that a further dispute, which the later account of meaning must answer. Nor do I follow Brentano in treating every mental phenomenon as conscious; directedness alone will not decide that question in this book.

> Return to the wanting, or try an anger that concerns nothing and a perception that presents nothing. Vague does not mean undirected: you can want something without knowing quite what, or see a shape without recognizing it. The exercise asks whether, after removing everything the state concerns, you can still identify the desire, anger, or perception you began with. Directedness does not look like a coat those states merely put on.

Two apparent counterexamples deserve daylight: diffuse mood and bodily pain. Neither is obviously blank. A mood can present the world as inviting, flat, or threatening without singling

out an object; pain can present the body as damaged or requiring withdrawal. Chapter 7 tests that answer rather than taking it for granted. A wholly undirected mental state would defeat the necessity claim. An undirected feature within an otherwise directed state would instead challenge the stronger claim that directedness exhausts its character.

Anger-at-the-delay and anger-at-the-remark differ partly in what each concerns. Remove that difference and we lose one way of telling the states apart. This supports giving directedness a constitutive role; it does not yet show that every feature of either anger consists in directedness. Pain and mood must face that stronger test on their own terms.

The inner/outer story may begin causally with light reaching the eye, signals traveling through the nervous system, and an event occurring in here. None of that shows that the event already counts as a complete mental state before we say what it concerns. A complete neural description need not give a complete description of it as anger or perception. My objection concerns treating directedness as a later addition to an already complete mental state. An inner-object account can reject that proposal too.

The use I make of Brentano changes that explanatory order. A mental state does not first arise as a bare inner occurrence and acquire an object afterward. It exists as a directed state: anger-at-the-delay, fear-of-the-dog, seeing-the-tree. The object need not exist, but the directedness belongs to the state from the start. This is the proposal the familiar examples support and the harder ones must test.

### 1.3 Whose Aboutness? Original and Derived

One qualification remains: not everything that points counts as mental.

A sentence concerns the world; a photograph, the wedding it shows; a road sign, the curve ahead. If aboutness marks the mental, has the sign become a mind?

No. The difference turns on dependence upon an interpreting practice.[2] Drivers make the symbols mean the curve by reading them so; left alone, paint remains on metal, legible to us and not to the rain.

Strip away every practice that reads the sign and its meaning vanishes. Philosophers call this *derived* intentionality: aboutness on loan, real in use but dependent on interpreters.

A state with *original* content works differently. Its own world-involving history and use help fix what would make it accurate or mistaken; an interpreting practice does not exhaust that standard. *Original* does not mean uncaused, uninherited, or self-created. Evolution, education, culture, engineering, and learning may all stand in the state's pedigree. The contrast concerns what fixes correctness, not whether the state sprang from nowhere. The road sign carries real but derived content. Neither having meaning nor carrying a useful signal, by itself, makes the sign a mind. Chapter 14 develops that distinction.

Whether borrowed meaning can run all the way down with no interpretation-independent content remains contested. Daniel Dennett rejects a sharp original/derived divide. On his interpretationist account, intentional descriptions earn literal standing when they capture real patterns

whose predictive and explanatory success is not merely in the interpreter's eye. This book retains the distinction as a diagnosis of semantic dependence, not a gate through which all genuine meaning or mindedness must pass. Dependence can vary by content and grain: a learner may inherit a word from a practice while acquiring new, world-answerable uses of it.

### 1.4 The Sealed Room as Inheritance

Put the argument together and the mind looks world-directed from the start: toward objects, fields, states of affairs that may or may not obtain, and sometimes no more than how things stand, with nothing singled out. This does not yet prove externalism or exclude every inner-object account. Brentano's own account combines directed acts with objects within them. I begin with directedness as a feature any account must explain; whether awareness reaches the world through an inner object remains open. Why, then, does so much of philosophy, science, and ordinary talk proceed as though we sit behind our eyes watching a private show?

Because the inner theater has a history. The picture that puts an inner object between the perceiver and the world grew through influential readings of Descartes and his successors. Later scientific talk often kept the arrangement while replacing ideas with brain states. A genealogy will not show the room false. It can show that we were never obliged to begin inside it, and hand the burden back to anyone who asks us to stay.

### Chapter Summary

**The claim.** Intentionality names the directedness Brentano proposed as the mark of the mental. The book takes that proposal into a different framework: mental states come already concerned with something, without requiring an object lodged within them.

**Where the argument stands.** The familiar examples support a constitutive role for directedness, not yet the stronger claim that it exhausts every mental feature. The road sign shows why meaning alone cannot establish a mind. Original content requires some independence from interpreting practice, not independence from history or inheritance.

**What's left open.** Whether diffuse moods and bodily pain are wholly directed, whether basic directedness requires satisfaction-conditioned content, whether a machine's mechanisms earn such content, and whether the larger system understands by means of it wait for later argument. None is settled here.

**The hand-off.** A mind world-directed through and through makes the sealed chamber look less like a discovery than a room we were handed. So the next chapter steps inside it to watch the walls go up, brick by brick, and then asks what the finished room can say in its own defense. What it says best, it says as a question — the one that makes the picture look unavoidable.

### Footnotes

**1.** Franz Brentano, *Psychology from an Empirical Standpoint*, trans. Antos C. Rancurello, D. B. Terrell, and Linda L. McAlister (London: Routledge, 1995 [1874]), Book II, ch. 1, p. 88. The canonical formulation: "Every mental phenomenon is characterized by what the Scholastics of the Middle Ages called the intentional (or mental) inexistence of an object, and what we might call, though not wholly unambiguously, reference to a content, direction toward an object (which is not to be

understood here as meaning a thing), or immanent objectivity." The Scholastic lineage matters: intentionality's roots run through Aquinas and Aristotle's account of the intellect's receiving the form of an object without its matter. Brentano's innovation was to elevate intentionality from one property of the mental among others to its *defining* criterion — the mark distinguishing mental from physical phenomena. For the framing of the thesis as the mark of the mental and a defense of its sufficiency as well as its necessity, see Tim Crane, "Intentionality as the Mark of the Mental," in *Contemporary Issues in the Philosophy of Mind*, ed. Anthony O'Hear (Cambridge: Cambridge University Press, 1998), 229–251 (Royal Institute of Philosophy Supplement 43). Crane's own thesis is the weak one — that all mental states are intentional "regardless of whether these states also have non-intentional properties" — so the claim this chapter makes sits inside his, not beyond it. He stops short of the identity claim the book argues later, assigning that to "contemporary intentionalists like Michael Tye and Gilbert Harman," and he names "the naggingness of a toothache" as a property of pain he takes to be wholly non-intentional (Crane, "Intentionality," *Routledge Encyclopedia of Philosophy*, 1998, sec. 3). The ally who supplies the mark does not follow the argument all the way, and the case where he stops is the case Appendix B.6 and Chapter 7 take up. Daniel D. Hutto and Erik Myin make the prior challenge explicit in *Radicalizing Enactivism: Basic Minds without Content* (Cambridge, MA: MIT Press, 2013), esp. the preface and chs. 1, 4, and 6: intentional directedness and even perceptual experience may occur without specified conditions of satisfaction. The book therefore cannot define their rival away by using *directedness* and *content* interchangeably. Note that "intentional inexistence" in Brentano's usage names the object's *existing-in* the mental act (Latin *in-existere*, on the model of Aristotelian immanent form), not the object's nonexistence; the standard English rendering has misled generations of readers, a confusion traced and corrected in Crane, "Brentano on Intentionality," in *The Routledge Handbook of Franz Brentano and the Brentano School*, ed. Uriah Kriegel (London: Routledge, 2017), 41–48. On the distinction between content and object, see Kazimierz Twardowski, *On the Content and Object of Presentations* (1894), and Husserl, *Logical Investigations*, Fifth Investigation, sec. 11. Husserl denies that the intentional object is a real constituent of the act, whether or not that object exists; his example of Jupiter does not relocate a nonexistent object into the external world (quoted and discussed in Crane, "Brentano on Intentionality"). The book carries directedness into its own world-involving framework rather than attributing that whole framework to Brentano or Husserl.

**2.** John Searle gives the intrinsic/derived vocabulary an influential modern formulation in *Intentionality: An Essay in the Philosophy of Mind* (Cambridge: Cambridge University Press, 1983), ch. 1; he sharpens the observer-relative framing in *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992). The underlying problem reaches well beyond his account: how public signs inherit meaning and how thoughts or mechanisms acquire content that does not depend on further interpretation. For the analysis of conventional meaning through speaker meaning, see H. P. Grice, "Meaning," *The Philosophical Review* 66, no. 3 (1957): 377–388. For naturalistic programs aimed at underived content, see Fred Dretske, *Knowledge and the Flow of Information* (Cambridge, MA: MIT Press, 1981); Ruth Garrett Millikan, *Language, Thought, and Other Biological Categories* (Cambridge, MA: MIT Press, 1984); and Nicholas Shea, *Representation in Cognitive Science* (Oxford: Oxford University Press, 2018). The present chapter uses *derived* and *original* to mark dependence on an interpreting practice, not allegiance to a specifically Searlean theory, and it does not reserve

underived content for whole biological minds. Selection, learning, and producer-consumer organization may establish original mechanism-level content. Part Four therefore asks separately whether machine mechanisms possess such content and whether the larger system believes, understands, or acts by means of it. For the interpretationist alternative, which rejects any sharp distinction between original and derived intentionality, see Daniel Dennett, *The Intentional Stance* (Cambridge, MA: MIT Press, 1987).

## Chapter 2: How the Inner Theater Was Built

> *"Nothing is ever really present to the mind, besides its own perceptions."*
>
> — David Hume, *A Treatise of Human Nature*

---

CHAPTER OVERVIEW

Most of us inherit the inner theater as if someone found it inside the skull. I trace how it took shape through Descartes' search for sure knowledge, later accounts of how ideas arise, and the ways those views were read. Talk of brain models, stored memories, and machine memory can likewise turn real work in the brain into a thing that someone must inspect. The picture then makes color, misleading sight, and the feel of a private inner life look like proof of a show behind the eyes. Some ways of stating these puzzles already assume the room they claim to find, though the questions can survive without it. History cannot refute the theater, but it can show why we need an argument before we agree to live inside it.

### 2.1 The Cartesian Tradition

René Descartes wanted to build knowledge on a foundation that could not crack. So he went looking for something he could not doubt, and found that almost everything cracked. The senses deceive: he had been fooled before, and what proof did he have that he wasn't being fooled right now? Even arithmetic felt shaky. Perhaps some sly demon was tampering with his mind, so that two and two only *seemed* to make four. One by one he threw the candidates out, until one thing was left standing — the doubting itself. He could not doubt that he was doubting. *Cogito ergo sum.* I think, therefore I am — though the tag itself belongs to the *Discourse* and the *Principles*; the *Meditations*, where the doubt is staged most starkly, reaches it as "I am, I exist." Everything else might be an illusion; the thinker having the illusion could not be.

Notice what that proves, and what it doesn't. It proves there is a thinker. It proves nothing about the world outside. Descartes has planted certainty *inside* — in the act of thinking, in the mind's view of its own contents. The world beyond stays a guess. Solid ground lies only within the mind.

The next move proved fateful in the reading of Descartes that became influential: *ideas* supply what the mind encounters immediately, and the world must be reached through them.[12] Look at a red apple. On this reading, awareness first meets your idea of it; the apple stays one step away, on the far side of the representation. Descartes also permits a different reading, on which an idea names a mode of thinking rather than an item inspected in the apple's place. That difference matters. The historical subject here is the veil-reading and its influence, not an uncontested verdict that Descartes put a colored picture inside the head.

Descartes noticed further that an idea of a thing need not *resemble* what causes it. In the *Optics* he resisted the demand that representation work through resemblance. The word "apple" need not look like an apple to concern one; neural activity need not look like its object either. This insight helps free representation from picturing. It also leaves a question for any account, including ours:

what makes a state represent one thing rather than another? The veil-reading adds a separate difficulty. If all you meet directly are your own ideas, how could you step outside them to check their relation to the world?

Descartes did not intend to leave the world a permanent guess. The *Meditations* attempt to recover knowledge after radical doubt, and a nondeceiving God supplies the guarantee. That recovery faces the famous charge of a circle: can clear and distinct reasoning establish the God whose guarantee secures such reasoning? Descartes had replies; the charge is not a refutation simply because it can be put in a sentence. The relevant pressure here concerns the veil-reading: once private ideas supply the whole starting evidence, the relation to an independent world needs a defense that cannot simply assume the contact placed in doubt.

Berkeley would reject mind-independent material substance while retaining sensible objects and the spirits that perceive them. Hume would press the limits of our rational warrant for an independently existing world. Their answers differ sharply. Both expose the difficulty of beginning with ideas and then asking how those ideas reach beyond themselves.

### 2.2 The Empiricist Renovation

The British Empiricists gave experience a central job: explain how our ideas and knowledge arise. That project need not begin with a private theater. But on the reading that came to dominate their reception, experience supplied ideas within the mind, leaving the world that caused them outside. Each thinker tried to resolve a difficulty inherited from that arrangement.

John Locke (a physician by training, who spent the better part of two decades on his great Essay before letting it out into the world) was read as keeping the core Cartesian picture, and it is that reading which traveled: what we meet directly are our own ideas, not the world. What he changed was where ideas come from. Not from innate ideas or principles, he held, but from experience. Every idea traces back to one of two springs: sensation, what comes in through the senses, or reflection, what the mind notices about its own workings. The newborn mind is a blank slate. Experience does the writing.

For perception, that meant the senses hand us *simple ideas*, raw bits like color, warmth, hardness, which the mind builds into ideas of whole objects. Sit by a fire and you take in simple ideas of heat, light, and color; your mind assembles them into the idea of the fire. On the veil-reading, the fire itself, the thing in the grate, remains beyond immediate awareness. What you meet is the assembly your mind has run up.

To keep this from sliding into full idealism, Locke split an object's qualities in two. *Primary* qualities (shape, size, solidity, motion) really belong to the thing, independent of us. *Secondary* qualities (color, taste, smell, warmth) work differently. They are *powers* in the object to produce sensations in us, and the sensation a power produces resembles nothing in the object, unlike our idea of its shape, which does.[3] So when you see the red of an apple, you are not taking in a color the apple owns the way it owns its shape. You are registering a sensory idea the apple's powers stir up in you. Those powers belong to the apple; the idea belongs to your experience.

Locke reached for scientific respectability. Physics describes the world in mass and motion, not in redness and sweetness, and he wanted his account to match. But the bill is steep. If colors and warmths as we experience them resemble nothing in the objects, then those objects hold powers to affect us but nothing resembling the sensory ideas they produce. The veil-reading makes those ideas what we directly encounter. The world thins out; the inner life swells; Descartes' gap widens.[4]

George Berkeley spotted a crack in Locke's split. He came to the problem as a young fellow of Trinity College Dublin, a clergyman set on defending religion against the materialism creeping through the new science, and the crack he found would serve that end. (The bishopric, at Cloyne, came decades later; the *Principles* is the work of a philosopher barely twenty-five.) If all we ever meet directly are our own ideas, how could we check which ideas resemble the world and which don't? We would have to step outside our ideas and set them beside the things themselves — the one thing we can never do. So the whole notion of a primary quality "resembling" something mind-independent falls apart on Locke's own terms. Berkeley's conclusion was bracing: for sensible objects, *esse est percipi*, to be is to be perceived. He denied a material substance behind the ideas, not the minds perceiving them. Sensible objects consist of ideas; spirits perceive and act. For Berkeley this meant not despair but rescue: the familiar world remains, upheld by a divine mind.

Berkeley insisted he was denying nothing the plain person believes: the cherry stays as red, as round and as edible as it ever was, and only the philosophers' inert matter behind it gets struck out. Grant him that and the argument offers a redescription, not an elimination. His appeal to God also gives the sensible world an order independent of any finite person's wishes. I do not adopt his idealism; this book starts from a mind-independent physical world. But invoking divine order does not refute him, and ridicule cannot settle the metaphysics. His challenge remains: if ideas alone supply our evidence, what justifies the extra material world the realist places behind them?

David Hume pressed hardest on the difference between rational justification and the beliefs we cannot help having. His skeptical arguments question our grounds for continued unperceived objects, necessary causal connections, and a self beneath passing perceptions. Yet he did not expect anyone to stop believing in bodies. Nature, he thought, had not left that to our choice. The difficulty lay in reconciling ordinary confidence with what philosophical reflection could warrant. Hume's wider work in morals, politics, and history belongs beside that inquiry, not as evidence that a defeated philosopher abandoned it.

Philosophers have a name for what all this builds: the *veil of perception*.[5] On the Cartesian-Lockean picture, we never perceive the world directly. We perceive our representations of it: our ideas, our inner images. The world sits behind the veil, causing the show but never stepping onto the stage. And granting the causal story doesn't help. Yes, the red you see traces back, along a chain of events, to a surface that reflects light a certain way — but you are still stuck with the gap. The surface is one thing. Your experience of its redness is another. How the two ever meet stays a mystery.

A word on those inner *images*, lest the target look like a cartoon. The careful sense-datum theorists hung no little colored pictures behind the eyes; Russell, Price, and Broad took pains to

deny that sense-data are mental images at all.[6] The thread this book pulls runs under the cartoon: the move that slips an *immediate object* of awareness (idea, datum, appearance, model) between the perceiver and the world, reached, if at all, only by inference. Whether that object is pictured as a picture is beside the point. The interposition is the target, and every version shares it.

### 2.3 How Neuroscience Inherited the Picture

Modern neuroscience might have dissolved the theater. Some descriptions instead preserve its floor plan. In them, the theater has simply been re-staffed with neurons.

Neuroscience tells a story of hidden perceptual machinery. Light falls on the retina, photoreceptors convert it into electrical signals, those signals travel through the optic nerve to the lateral geniculate nucleus, which relays them to the primary visual cortex, which projects to a cascade of higher visual areas that extract edges, orientations, depths, motions, colors, faces, objects. By the time consciousness enters the picture (if "enters" is even the right word), an enormous amount of processing has already happened, most of it below the level of any awareness. What we experience as the immediate, effortless perception of a scene turns out to be the output of a massively parallel computational enterprise running largely in the dark. The neuroscientist Anil Seth describes the brain as a "prediction machine," actively generating a "controlled hallucination" of the world rather than passively receiving it.[7]

Trouble enters with the description. When "the brain constructs a model" slides into "the subject perceives that model," the model's location inside the brain seems to place awareness there too. But that conclusion belongs to an inherited frame, not to the neural evidence.

The phenomenologist replies: you see the fire, not a model. The model theorist can answer that a transparent model presents itself as the fire. Both descriptions fit. Talk of models alone therefore cannot settle what awareness reaches.[8]

A charitable reading remains true. The brain processes information, predicts, resolves ambiguities, and recovers three-dimensional structure from retinal input. None of that shows that awareness lands on the processing rather than the world. Machinery may lie in the brain while its object does not.

The needed distinction separates the state that represents from what it presents. *The Seeing and the Seen* earns that tool after we inspect experience. For now: that the brain builds the seeing does not put the seen inside the skull.

Five claims crowd together in the inherited picture, and the rest of the book needs them kept apart.

1. Perception depends on causal mediation through the senses and brain. It does.
2. Neural systems form internal states that represent, predict, and control. They may.
3. We perceive the world only by first perceiving one of those states—an idea, datum, image, or model. That is the veil of perception.
4. The processing culminates at a central place where its results are presented for consciousness. That is the Cartesian Theater in Dennett's strict sense.
5. An observing self witnesses the presentation, perhaps accompanied by intrinsic private qualities disclosed only to it. Those are further claims again.

The first two do not entail the last three. Internal representation supplies machinery, not an audience; distributed neural work need never converge on a screen. This chapter targets the interposed object and the centralized presentation to a witness. Later chapters ask separately whether phenomenal character contains intrinsic mental paint and how an integrated subject emerges from the distributed organization. A theory can lose either dispute without making neural representation itself Cartesian.

### 2.4 The Same Move in Memory, and in Machine "Memory"

Once you can see the inner-theater move in perception, you start to find it everywhere. Memory makes the cleanest second case.

> Picture the kitchen you grew up in. Perhaps it comes as a faded reel running somewhere behind the eyes, colors a little washed out, the kitchen viewed from a vantage that no longer exists. That can make remembering sound like watching an inner film. But the report alone does not settle what, if anything, gets watched.

Keep two things apart that the exercise runs together. Summoning imagery is not the same act as remembering: you can call up a vivid kitchen you never stood in, and you can remember a great deal about the one you did without any picture arriving at all. The inner reel, strictly, is a story about *imagery*. The same inference gets made about memory itself, and needs its own defense.

The theater-reading supplies a ready account: the mind contains a chamber, perception leaves traces in it, and some inner agency calls those traces back and watches them. That picture can feel true. When you summon the kitchen, it may seem to be a thing summoned, an inner item produced for inspection.

The faded colors and vanished vantage point tell us how the kitchen comes to mind. They do not by themselves reveal an inner picture produced for inspection. Remembering concerns the absent kitchen; whether it does so through an immediately apprehended image needs a further argument.[9] The distinction we need in perception returns here: what makes remembering possible need not be what we remember. For now, the kitchen exercise raises that question rather than settling it.

Notice how far the cinema picture reaches. Contemporary discourse around artificial intelligence has quietly inherited it, and Part Four will develop what follows from that. The point available here is narrower, and it is enough: if we reduce human memory to storage and retrieval, we choose the terms of comparison before asking what remembering involves. A database might excel on capacity or accuracy without settling anything about recollective experience. Nor does storing information itself imply an inner spectator.

### 2.5 Five Questions, One Picture

Once the history is visible, return to the questions that make the inner theater feel unavoidable. They no longer arrive as neutral data. They arrive inside a picture with a history, and each carries an answer that springs up so readily it scarcely seems like a theory at all.

1. Start with the redness of the tomato. Does it belong out there on the fruit, or does the brain add it? The answer that springs to mind: the brain adds it. Out in the world, on the modern scientific view, we find only invisible wavelengths reflecting off colorless surfaces; the warm red you actually see must come from somewhere inside the brain or mind, or so it seems.
2. Push a straight stick halfway into water and it looks bent. You know better — pull it out and the stick stands as straight as ever. So what is the bent thing you plainly see? The reflex says you see a bent *something*; and since the stick itself stands straight, that bent something cannot belong to the stick. It can only show up as an appearance, standing in for the stick, and the bent appearance belongs to you, in here, or so it seems.
3. When Shakespeare's Macbeth, in the grip of hallucination, reaches for a dagger that does not exist, what does he reach toward? Something hangs in the air for him (dagger-shaped, edged, *seen*) and yet the air holds nothing. The dagger must be an item in his mind that he sees, or so it seems.
4. When you and a friend both point at the tomato and say "red," do you even share the same red experience?[10] The natural thought: no one could ever really know, since each of us has access only to our own case. Where you see red, I might see green — and since we would both go on calling ripe tomatoes red, no test would catch it. So the experience turns private: an inner screening with an audience of one, so it seems.
5. And when a machine answers you in fluent, feeling sentences, does anyone live behind the words? The reflex that takes the question seriously says you cannot tell from outside. Either some inner light burns, some flicker of what it is like to be the machine, or nothing whatever inhabits the inside, and the fluent words amount to a very good trick. Whatever settles the matter lies behind the words, out of view, or so it seems.

These formulations divide into two families. Vanished color, bent shape, and the dagger in the air invite the same inference: not in the world, therefore in here. The private spectrum and the machine's hidden light instead borrow the theater's imagery. That imagery cannot serve as evidence for the picture it already assumes. The underlying questions survive without it, however. Could two experiences differ in feel despite matching in other respects? What would make an artificial system conscious? Neither question requires a screen or spectator, and the later chapters must answer both without treating the questioner as a concealed Cartesian.

Put all five together and you get a single picture so familiar most of us never notice we hold it. According to that picture, behind the eyes lives an inner theater. The *real* you sits in the audience. When you see the tomato on the counter (round, waxy, that red that seems to glow from inside), you see only your inner representation of it. The fruit itself you only infer.

### 2.6 The Picture's Credentials

Do not wave the inner theater away. The picture has earned a wide following. Neuroscience tells us, correctly, that the brain builds models: the redness you see emerges at the end of a long process from photon to retina to cortex, and the process bends to fooling, lesioning, drugs, rewiring. Close your eyes and the tomato disappears; open them and it returns. Seeing looks separable from the

seen — something that could in principle run without it. And above all, the picture preserves the one phenomenon any truthful description has to keep on the table: that there *is* something it is like to see red, a felt quality present to you and to no one else, as private as a held breath. Put those together and the inner theater stops looking like a theory. It looks like what taking science and your own experience seriously *forces* you to accept.

Follow the picture further and the troubles begin. If you only ever perceive your inner image, how do you ever *know* the tomato exists? And where does the redness live if not on the tomato, not in the photons, and not in the firing neurons? Does it live in some third place belonging to neither world nor brain?

Pour a cup of coffee and take the first sip of the morning. Something happens there that a physicist could describe at enormous length. Volatile compounds reach receptors in the nose; dissolved acids and bitter alkaloids meet the tongue; heat sensors report; signal after signal runs inward and fans across the cortex, where memory and expectation arrive to help. Every step of that account may be true. One day it may even be finished. Read it back to yourself anyway, and something seems to have gone missing from it. It says nothing about the warmth, or the faint bitterness, or the way the roast announces itself before you have decided anything at all. Why should any of that processing feel like anything from the inside?

The question has a name. So does the paper that gave it one: David Chalmers published "Facing Up to the Problem of Consciousness" in 1995,[11] and thirty years on, the question has not gone away. He gives the name *easy problems* to the ordinary business of explaining what a brain does: how it discriminates, integrates, attends, reports. Easy not because anyone can solve them, but because we know what solving them would look like. The hard problem names what remains. Grant every easy problem its answer. Grant the entire functional story, start to finish. It still seems to make sense to ask why the lights are on: why all that processing does not simply run in the dark.

This chapter will not answer any of it. Chapters 9 and 11 take the question up in earnest.

### 2.7 Indirect Realism and the Invisible Experience

The interposed-object version of that picture gives one familiar form of *indirect realism*. It raises a question: when you see the tomato, does reflection reveal a second thing through which you see it? The indirect realist may reply that the intermediary presents itself as the tomato, not as a mental item. Nor must the act of seeing itself appear as something visible. Chapter 4 asks what this transparency shows, and what it leaves open.

### 2.8 Why the Picture Persists

The inner-theater picture has been under serious philosophical pressure for a long time. And yet it persists — not just in philosophy departments, but in popular neuroscience, in AI discourse, in the way people talk about their own minds at dinner tables. Why?

Part of the answer is grammatical. "I *have* an experience" and "I *feel* a pain" invite an inner object one possesses. But "I see red" need not describe an encounter with inner redness; it may describe a perceptual relation to a surface.

Part of the answer is scientific prestige. When neuroscience says the brain processes signals and constructs models, and when this story appears in authoritative textbooks and popular books by distinguished researchers, the inner-theater picture gets wrapped in the authority of empirical science. It becomes harder to question because it seems to have been confirmed rather than merely assumed.

The picture also takes the right things seriously before going wrong in the explanation. It is a good first guess that has overstayed its welcome. The plain sense of direct contact with the world remains the incumbent; a theory that replaces it with a veil carries the burden of proof.

### 2.9 The Inner Theater, in Place

The inner theater does not stand on history and habit alone. Its strongest argument begins with the bent stick: if it looks bent and is not, perhaps what you directly perceive cannot be the stick. The next chapter tests that inference.

### Chapter Summary

**The claim.** The inner theater records no discovery. It grew out of a historically contingent philosophical inheritance: put certainty inside and place the world outside. It then acquired furniture from empiricism and borrowed authority from neuroscience.

**Where the argument stands.** An influential reading of Descartes and Locke placed ideas between perceiver and world; Berkeley and Hume pressed the difficulties of that starting point in different ways. Modern neural facts show causal mediation, model-building, and reconstruction, not that awareness stops at an inner object. Grammar and scientific prestige help explain the picture's appeal without proving it true.

**What's left open.** Genealogy removes inevitability; it does not refute the theater. Illusion, hallucination, privacy, and the hard problem still deserve direct tests.

**The hand-off.** The first test begins with the bent stick. If appearance can mislead, must awareness terminate at a private image rather than the object itself?

### Footnotes

**1.** The gentler reading of Cartesian "ideas" — as acts or modes of thinking rather than tertia inspected in the world's place — was pressed by Descartes' own contemporaries; Arnauld is the locus classicus. For the competing representationalist and direct-realist interpretations, see Kurt Smith, "Descartes' Theory of Ideas," *Stanford Encyclopedia of Philosophy*, substantive revision September 18, 2025, sec. 1, https://plato.stanford.edu/entries/descartes-ideas/. Smith surveys the alternatives; this secondary account does not by itself establish the history of their transmission. For the parallel Locke case see note 4 below. Thomas Reid, *Essays on the Intellectual Powers of Man* (1785), is the source of the influential attribution of an "ideal theory" to Descartes, Locke, Berkeley, and Hume, and thus one of the reception's principal transmitters; the nineteenth- and twentieth-century textbook traditions carried it, and the label *veil of perception* itself dates only to Jonathan Bennett in 1971 (see note 5). Nicolas Malebranche's vision-in-God theory requires a separate distinction: intelligible ideas exist in God, whereas sensations modify the finite perceiver's mind. His ideas therefore cannot simply be classified as private inner objects within that mind. See

*Dialogues on Metaphysics and on Religion* (1688), Dialogue I, sec. 10, and Dialogue IV, sec. 10, which distinguish intelligible extension, material bodies, and sensations. A genealogy of the picture does not require Descartes guilty; it requires only that the picture got transmitted, and it plainly did.

**2.** On the Cartesian representational theory of perception, see especially René Descartes, *Meditations on First Philosophy*, trans. John Cottingham (Cambridge: Cambridge University Press, 1996 [1641]), Third and Sixth Meditations; and *Optics*, in *Discourse on Method, Optics, Geometry, and Meteorology*, trans. Paul J. Olscamp (Indianapolis: Hackett, 2001). Discourse IV of the *Optics* resists the assumption that the mind must contemplate resembling images transmitted to the brain. Signs and words can stimulate thought without resemblance; even pictorial representation depends on selective resemblance and systematic departures from it. This separates representation from copying rather than establishing that representation fails. The remaining content-fixing question concerns what makes the relevant activity represent one object or property rather than another. The Cartesian-circle objection and Descartes' distinction between present clear perception and remembered demonstration appear in the Fourth Objections and Replies. See Gary Hatfield, "The Cognitive Faculties," in *The Cambridge History of Seventeenth-Century Philosophy*, ed. Daniel Garber and Michael Ayers (Cambridge: Cambridge University Press, 1998), for helpful historical context.

**3.** John Locke, *An Essay Concerning Human Understanding*, ed. Peter H. Nidditch (Oxford: Clarendon Press, 1975 [1689]), Book II, ch. viii, esp. secs. 9–10 and 15. Locke's distinction between primary and secondary qualities has generated an enormous secondary literature. For the purposes of this book, the important point is its role in driving the inner-theater picture: secondary qualities, understood as powers, remain in objects; their sensory effects are ideas in perceivers. Locke denies that these ideas resemble the powers producing them. On the veil-reading, making the ideas immediate objects of awareness widens the gulf between the world as physics describes it and the world as it appears. Jonathan Bennett, *Locke, Berkeley, Hume: Central Themes* (Oxford: Clarendon Press, 1971), remains a useful critical survey of the positions and their internal tensions.

**4.** The reading that resists treating Locke's "ideas" as a veil of interposed objects is developed by John W. Yolton, *Perceptual Acquaintance: From Descartes to Reid* (Minneapolis: University of Minnesota Press, 1984), and *Perception & Reality: A History from Descartes to Kant* (Ithaca: Cornell University Press, 1996). On Yolton's account Lockean ideas are acts or contents of perceiving — the *having* of which is the perceiving of the world — rather than tertia perceived in the world's place. Michael Ayers, *Locke*, 2 vols. (London: Routledge, 1991), esp. vol. 1, *Epistemology*, should not be assimilated to that reading without qualification: Ayers resists the crude veil picture but finds in Locke an unresolved ambiguity, and titles a chapter of his own "Ideas as intentional acts and ideas as intentional objects." The contested state of the exegesis is itself part of this section's point. The contrary "veil of ideas" reading is the one entrenched in the textbook tradition and in the scientific vocabulary descended from it (see, e.g., Jonathan Bennett, *Locke, Berkeley, Hume: Central Themes* [Oxford: Clarendon Press, 1971]). That is why the book engages the veil-reading: not as a verdict on Locke exegesis, but as the picture that left the study for the laboratory. Whoever the historical Locke was, the interpretation his *Essay* set loose is the one this chapter traces. Its private sensory effects will return later, in newer dress, as qualia.

**5.** The "veil of perception" label entered the literature with Jonathan Bennett, *Locke, Berkeley, Hume: Central Themes* (Oxford: Clarendon Press, 1971), esp. 69, and is standardly credited to him; its entrenchment as the standard shorthand for the epistemological predicament generated by indirect realism owes much to Howard Robinson, *Perception* (London: Routledge, 1994). It should be distinguished from the "veil of ideas" discussed in Locke scholarship, which refers to the mediating role of mental representations specifically in the context of Locke's theory — the broader "veil of perception" captures the general predicament of any view that makes inner representations the immediate objects of awareness. For a direct critique, see Tim Crane, "Is There a Perceptual Relation?", in *Perceptual Experience*, ed. Tamar Szabó Gendler and John Hawthorne (Oxford: Oxford University Press, 2006).

**6.** That sense-data need not be construed as mental images, and were not so construed by the most careful sense-datum theorists, is explicit in the founding statements of the view. Bertrand Russell takes sense-data to be objects of acquaintance whose status as mental or physical he holds open (*The Problems of Philosophy* [London: Williams and Norgate, 1912], chs. 1–2), and H. H. Price (*Perception* [London: Methuen, 1932]) and C. D. Broad (*Scientific Thought* [London: Kegan Paul, Trench, Trübner, 1923]) both insist that a sensum is not an image inspected in the head. What the chapter targets is therefore not the literal-picture caricature but the structural commitment common to all versions — the positing of an immediate, non-worldly object of awareness from which the external world is reached only by inference. The classic exposure of that structural move (as opposed to the cartoon) is J. L. Austin, *Sense and Sensibilia*, reconstructed by G. J. Warnock (Oxford: Oxford University Press, 1962), taken up in the next chapter.

**7.** Anil Seth, *Being You: A New Science of Consciousness* (London: Faber & Faber, 2021), esp. chs. 4–6. Seth's "controlled hallucination" formulation is a deliberately provocative way of expressing the predictive-processing framework associated with Karl Friston and, in earlier form, Hermann von Helmholtz's account of perception as "unconscious inference." The claim is not that ordinary perception is pathological but that the generative process underlying hallucination and veridical perception is the same — a point that feeds directly into the argument from hallucination taken up in the next chapter. For a more cautious philosophical treatment of what the predictive-processing framework does and does not entail about the objects of awareness, see Jakob Hohwy, *The Predictive Mind* (Oxford: Oxford University Press, 2013). Bayesian versions of the framework, which model perception as approximate probabilistic inference, are compatible with this book's account when read as claims about how subpersonal processing works rather than about what experience presents. Ned Block presses that gap, asking why conscious perception does not seem probabilistic if perceptual representation is, and argues that the best Bayesian approach does not require probabilistic representation at all ("If Perception Is Probabilistic, Why Does It Not Seem Probabilistic?," *Philosophical Transactions of the Royal Society B* 373, no. 1755 [2018]: 20170341). The present book takes no settled position on the predictive-processing framework as neuroscience; it uses it to illustrate how naturally the Cartesian-Lockean picture maps onto the vocabulary of contemporary cognitive science — a mapping that any adequate philosophy of perception must confront.

**8.** Causal mediation through brain processes does not entail that the immediate object of perceptual awareness is a brain state or model rather than the world. Galen Strawson, "Real Direct Realism: Reflections on Perception," in *Phenomenal Qualities: Sense, Perception, and Consciousness*, ed. Paul Coates and Sam Coleman (Oxford: Oxford University Press, 2015), 214–253, esp. secs. 1 and 9, defends this compatibility of representation and direct perception. A related distinction between phenomenologically available content and subpersonal representational content appears in Tim Crane, "The Intentional Structure of Consciousness," in *Consciousness: New Philosophical Perspectives*, ed. Quentin Smith and Aleksandar Jokic (Oxford: Oxford University Press, 2003), 33–56, sec. 5. Neither distinction by itself establishes the historical transmission of Cartesian assumptions into neuroscience.

**9.** For the distinction between remembering a fact and the imagery or feelings that may accompany it, see Michael Tye, *Ten Problems of Consciousness* (Cambridge, MA: MIT Press, 1995), ch. 1, esp. 4–5, where he observes that "in and of itself there is nothing it is like to remember that September 2 is the date on which I first fell in love." (The transparency material in *Consciousness, Color, and Content* [Cambridge, MA: MIT Press, 2000] is ch. 3, not ch. 1; ch. 1 there is the knowledge-argument essay.) The distinction between *episodic* memory (the recall of a participated event, with a recoverable point of view) and *semantic* memory (the storehouse of facts, including facts about events never witnessed) traces to Endel Tulving, *Elements of Episodic Memory* (Oxford: Clarendon Press, 1983), who marks the asymmetry as one of *recollective experience* — autonoetic in the episodic case, noetic in the semantic — rather than as a flat presence or absence of phenomenal character; Tim Crane, *The Mechanical Mind*, 2nd ed. (London: Routledge, 2003), treats where the line falls as an empirical question common sense cannot settle. The main text needs only the weaker distinction between remembering and its accompanying imagery. Tye's quoted example supports that distinction, not by itself a general representational account of recollective experience. Whether remembering involves an immediately apprehended inner image requires a separate argument; the kitchen exercise does not establish its absence.

**10.** The question raised here in plain terms — whether two people who agree in all their color words might nonetheless differ in the experiences those words track — is the inverted-spectrum hypothesis. The thought traces to Locke, who entertained the possibility that "the same Object should produce in several Men's Minds different *Ideas* at the same time" (*An Essay Concerning Human Understanding*, ed. Peter H. Nidditch [Oxford: Clarendon, 1975], II.xxxii.15). Sydney Shoemaker ("The Inverted Spectrum," *Journal of Philosophy* 79, no. 7 [1982]: 357–81) and Ned Block ("Inverted Earth," *Philosophical Perspectives* 4 [1990]: 53–79) sharpened the thought into the canonical challenge to functionalism: if spectrum inversion is coherent while leaving all functional organization intact, then functional role cannot exhaust phenomenal character. David Chalmers treats inversion as one of his central cases for the explanatory gap ("Facing Up to the Problem of Consciousness," *Journal of Consciousness Studies* 2, no. 3 [1995]: 200–219). Chapter 6 develops the strong-representationalist reply without claiming that ordinary color-space asymmetries prove every inversion impossible. The identity claim rules out phenomenal inversion with the complete character-fixing intentional profile unchanged; whether a proposed case can preserve that profile is the disputed step. Matching color words or overt behavior alone

does not establish the match. The main text deliberately states the puzzle in pre-theoretical form; the apparatus arrives later.

**11.** David J. Chalmers, "Facing Up to the Problem of Consciousness," *Journal of Consciousness Studies* 2, no. 3 (1995): 200–219, esp. 200–203. Chalmers distinguishes problems susceptible to functional or neural explanation—discrimination, integration, access, report, attention, and behavioral control—from the further question why any such processing comes with experience. The main text states that explanatory demand before later chapters dispute Chalmers's inference from the apparent gap to an ontological remainder.

# Chapter 3: The Argument from Illusion

> *"When I look at a straight stick, which is refracted in water and so appears crooked, my experience is qualitatively the same as if I were looking at a stick that really was crooked."*
>
> — A. J. Ayer, *The Foundations of Empirical Knowledge*

CHAPTER OVERVIEW

A straight stick looks bent in water, and the old argument asks where the bend could be if the stick stays straight. I give that question its strongest form, because careful people went behind the veil for reasons worth taking seriously. The case first puts a bent inner thing into illusion, then spreads that thing across all sight; hallucination adds a whole scene with no object to match it. Yet the same felt look or the same brain cause does not by itself place a screen between mind and world. Austin finds the slide from "looks bent" to "awareness of something bent," while the disjunctivist reminds us that two cases may seem alike without sharing one basic kind. The result gives us room for an alternative: illusion and hallucination do not force an inner object, though sight's vivid presence still needs explaining.

## 3.1 The Bent Stick: The Argument Stated

Picture a straight stick resting halfway in a glass of water. At the waterline it looks bent. You know the stick has not changed shape, but knowing that does not straighten the look. Now ask the awkward question: what are you directly aware of? Try answering before giving the bend a place to live.

The argument from illusion offers a compelling answer. Since the physical stick remains straight, something else must actually bear the apparent bend. That *must* introduces the extra premise. Philosophers later gave it a useful name, the **Phenomenal Principle**: whenever something appears to have a sensible quality, awareness must include something that really has that quality. Grant the principle and the first inner item has already entered the room. Call it a sense-datum, an appearance, an inner image. It bears the bend the stick lacks, and awareness lands on it rather than on the physical stick.[1]

That gives the argument its first joint. The second spreads the result from illusion to perception as a whole. Grant, for the moment, that the bent-looking case involves an inner item. A veridical perception of a straight stick and an illusory perception of a bent one can shade into each other continuously: lift the stick slowly from the water and watch the bend ease out. No point in that progression announces a switch in the *kind* of thing you confront. Ayer presents the inference to a common kind across the series, then questions whether it proves a matter of fact.[2] If awareness lands on an inner item at the distorted end, consistency seems to place an inner item at the veridical end too.

Keep the two joints apart.

1. The Phenomenal Principle moves from *the stick looks bent* to *something before awareness actually is bent.*
2. The continuous-series argument then moves from *an inner item appears in illusion* to *the same kind of item appears in every perception.*

A critic can reject either move. Deny the first and no inner object enters. Grant it but resist the second, and the veil theorist can ask where, along the smooth slope, awareness switches from inner to outer. That question gains force only after the first joint admits an inner object.

If both joints hold, the conclusion lands hard. In every perception, veridical or otherwise, the immediate object of awareness lies within. The world stands at one remove, behind the representation, on the far side of the glass.

It also carries a modern cargo. The conclusion gives human perceivers and text-only language models one problem in common: neither would reach a scene directly. The comparison stops there. Human perception comes embodied, causally governed by its surroundings, and shaped by a long history of successful action; a text-only model has no comparable perceptual loop. Still, if every mind meets only an inner display, one contrast Part Four needs has already disappeared. More than a stick bends on the outcome.

### 3.2 The Argument from Hallucination

A sharper version of the same argument presses from the limit case, where the world drops out entirely.

Actual hallucinations show how much a mind can present when no matching object occupies the scene. People with Charles Bonnet syndrome may see faces, animals, or patterned figures with nothing corresponding before them. A bereaved person may briefly see a late spouse in the kitchen doorway. These cases deserve care: people with Charles Bonnet syndrome often retain insight into the source of what they see, and bereavement experiences vary widely in force and interpretation. Clinical examples establish that vivid objectless presentation occurs. They do not establish the philosopher's stronger premise that hallucination can match perception perfectly from the subject's point of view.[3]

> The argument therefore asks us to consider a possible hallucination specified to meet that stronger condition: same felt character, same apparent scene, no external object, and no clue available from within the episode that anything has gone wrong. The subject does not feel himself generating an inner image. He feels himself *seeing*.

Two claims now do the work.

1. If a hallucination and a genuine perception can remain wholly indiscriminable from within, they must share the same fundamental mental kind. Call this the **common-factor premise**.

2. Whenever awareness presents an object and its qualities, something must exist that bears those qualities and serves as awareness's immediate object. Call this the **phenomenal-object premise**.

Since no external object exists in hallucination, the bearer must lie within; the common factor then carries that inner object into genuine perception.

Indiscriminability alone produces no sense-datum. An intentionalist can grant perception and hallucination a shared way of presenting the world without turning the presentation into an object one sees. Only the phenomenal-object premise brings an inner item into the common factor.[4]

Hallucination remains the sharper challenge. Illusion leaves a real stick available for misrepresentation; hallucination removes it. Anyone who rejects the inner item must explain detailed presentation without quietly reinstalling the scene inside the head.

Now watch the two pressures compound. Trace the causal story: the world reaches consciousness only after a long relay, crowded with processing, as the previous chapter described. Examine the phenomenology: the same apparent scene could, the argument claims, occur without any scene answering to it. The first pressure concerns causal mediation; the second concerns what experience can present. Neither by itself proves that awareness terminates within. Together they make the wall look solid. Science presses from one side, phenomenology from the other.

### 3.3 The Two Escape Routes: Austin and Disjunctivism

Two lines of resistance expose the disputed moves without yet settling the case.

J. L. Austin thought the whole argument ran on a grammatical sleight of hand.[5] He had spent the war in military intelligence and came back to Oxford with a plain habit: read the sentence people actually say. The sleight lives in a quiet trade: *the stick looks bent*, a remark on how perception presents its object, becomes *I am aware of something bent*, an existential claim about an inner item. Only the trade conjures that item into being. The conceptual order runs against it too: something must count as bent before we can say that it merely looks bent. Appearance-talk borrows its sense from object-talk and cannot supply the inner stuff from which the world gets rebuilt.

Disjunctivists press a different seam.[6] The argument from hallucination assumes that states indiscriminable from within must share a fundamental mental kind. Deny that, and the inference never starts: genuine perception may involve a relation in which the world itself figures, while hallucination belongs to another kind entirely. A convincing forgery can defeat inspection without becoming the original. Likewise, your inability to tell two experiences apart reports a limit in *you*; it does not settle what, beyond that limit, the experiences amount to.

> The hardest pressure on this route comes from a **causally matching hallucination**. Suppose an unusual intervention produces the same proximate neural conditions as a genuine perception, but without the perceived object. If a positive account of those neural conditions explains the hallucination's apparent scene and its effects, why does the same account not also explain the matching perception, leaving

> the world-involving relation with no work to do? At Berkeley, M. G. F. Martin built his most demanding defense of disjunctivism around this pressure. He refuses that positive common characterization: the matching hallucination gets described through its indiscriminability from perception, not through a shared representational nature. The cost is explanatory, not contradictory; the move from matching causes to an identical total mental state still needs an argument, and Martin refuses it.[7]

Austin and the disjunctivists therefore leave different inheritances. Austin identifies the illicit promotion of a way an object looks into a second object of awareness. *The Seeing and the Seen* completes his diagnosis by separating the vehicle of presentation from its object. Disjunctivism preserves another result the book will keep: indiscriminability never settles fundamental kind. The book parts with Martin over his negative account of causally matching hallucinations. We favor common representational content to explain the apparent scene, without making content another object seen. This does not require attributing a denial of every shared mental feature to disjunctivists generally.

An answer must explain objectless presentation, preserve the world's role in genuine perception, and find no private scene along the way.

### 3.4 What the Argument Would Cost Us

The arguments have not forced a veil upon us. Consider, nevertheless, what adopting one would cost. Those costs need to be weighed against the explanation it offers.

> Borrow a small case from epistemology. A man hallucinates a clock on the wall of his kitchen. The wall carries no clock; the space stands empty. But the hallucination is vivid and orderly, and at the moment he looks, the hallucinated clock reads 3:42. As it happens, the actual time is 3:42. He forms the belief that it is 3:42. The belief is true. He feels every bit as certain as he would with a real clock in front of him. Does he *know* what time it is?

No.[8] The time played no part in producing his true belief. Let the time advance, stop every working clock in the house, or carry him across a time zone; the hallucinated dial follows its own schedule. From within the episode, he cannot distinguish this state from one produced by a clock and functioning eyes. Outside that perspective, the difference looks decisive. One belief answers to the time. The other merely agrees with it.

The veil theorist owes you no apology for that kind of luck. Her inner display tracks the world along lawful pathways: move the tomato and the display changes with it. Reliable indirect perception would therefore differ from the hallucinated clock in exactly the respect the example exposes. The analogy isolates causal disconnection; it cannot show that a lawfully driven state

lacks perceptual contact. That leaves the real dispute: do we see the world by means of the state, or first see an inner bearer that depicts it?

One pressure concerns what perception consists in. If awareness terminates at an inner display, no improvement in fidelity removes the intermediary. Perfect correlation makes the display reliable; it does not remove it. A direct realist can therefore distinguish causal control from perceptual contact. The pressure disappears if content names the manner in which the tomato itself appears rather than a second object in its place. That possibility belongs to the next chapters.

The veil theorist replies that the world appears *as content*: an inner display *of* the tomato hardly leaves the tomato absent. The dispute turns on whether the world figures as the content of something seen or as what is seen. *The Seeing and the Seen* separates the vehicle of presentation from its object; *The Identity Claim* asks whether content needs any inner bearer with tomato-like qualities. Until then, the constitutive cost remains a charge to be tried, not a sentence already passed.

The second pressure concerns knowledge of the coupling. Every theory uses perception to investigate perception, so ordinary circularity proves nothing. Trouble begins only if the veil confines direct evidence to private inner items and requires an inference from them to the world. Retinal maps, psychophysics, and the stick pulled dripping from the water then arrive as further items inside the divide. Descartes secured the bridge through a divine guarantee. An epistemic externalist can refuse the demand: reliable coupling may yield knowledge even when the subject cannot prove it from within. The pressure therefore falls on the stricter veil account that treats private inner items as the whole evidential base.

Neither pressure refutes indirect realism. The veil theorist must explain why an inner object does not terminate awareness short of the world and how veiled evidence establishes a bridge beyond it. The direct realist must explain how mediated, fallible perception reaches the world without magic or infallibility. Nobody leaves the audit early.

Whatever perception turns out to involve, ordinary seeing presents itself as contact with the world, not acquaintance with a private stand-in. That suspicion, felt before anyone can prove it, supplies the thread the next chapters follow.

**Chapter Summary**

**The claim.** The argument from illusion needs two joints: the Phenomenal Principle places an inner bearer behind the bent appearance, and the continuous-series argument spreads that bearer across perception. Hallucination adds the possibility of an objectless but indiscriminable scene and a common-factor premise, but it reaches an inner object only by retaining the phenomenal-object premise.

**Where the argument stands.** Illusion and hallucination do not force an inner object: both arguments need a further premise about what presentation requires. The clock case isolates causal disconnection, not a defect shared by all indirect perception. Lawful coupling leaves the separate question of what awareness reaches.

**What's left open.** Austin's grammatical diagnosis awaits the vehicle/object distinction that completes it. The book keeps the disjunctivist insight that indiscriminability does not settle kind. Its preference for common content over Martin's negative account of causally matching hallucinations still owes a positive defense.

**The hand-off.** We need not accept the inner object, but we still owe an account of sensory presence without it. The next chapter begins with what experience presents when we attend.

**Footnotes**

**1.** Howard Robinson later labels the first inferential joint the Phenomenal Principle in *Perception* (London: Routledge, 1994), esp. 32: if something appears to a subject to have a sensible quality, then something of which the subject is aware must actually have that quality. The principle converts a manner of appearance into a further bearer of the apparent quality. J. L. Austin's criticism targets that conversion directly in *Sense and Sensibilia*, reconstructed from manuscript notes by G. J. Warnock (Oxford: Oxford University Press, 1962), esp. secs. III and VII.

**2.** The ancient pedigree concerns perceptual variation, not yet the modern inference to sense-data: Sextus Empiricus uses the oar that appears broken in water and straight outside it in *Outlines of Pyrrhonism* I.118–119. The epigraph appears in A. J. Ayer, *The Foundations of Empirical Knowledge* (London: Macmillan, 1940), 6. Ayer, 5–14, and H. H. Price, *Perception* (London: Methuen, 1932), supply classic expositions of the spreading argument. Ayer subsequently denies its factual conclusiveness: the assumptions about qualitative distinguishability and continuous variation can be denied without contradiction (12). Austin reconstructs the qualitative-similarity and continuous-series arguments in *Sense and Sensibilia*, 43–46. That progression supplies the second joint; it has force only after the first joint has placed an inner item in the illusory case.

**3.** The philosophical argument requires a *perfect* or subjectively indiscriminable hallucination, not merely a vivid clinical episode. Typical Charles Bonnet syndrome combines complex visual hallucinations with preserved cognition and, often, insight; see Dominic H. ffytche, "Visual Hallucinations and the Charles Bonnet Syndrome," *Current Psychiatry Reports* 7, no. 3 (2005): 168–179. Sensory and quasi-sensory experiences of deceased people also occur commonly in bereavement and vary in modality and interpretation; see Evelyn Elsaesser, Chris A. Roe, Callum E. Cooper, and David Lorimer, "The Phenomenology and Impact of Hallucinations Concerning the Deceased," *BJPsych Open* 7, no. 5 (2021): e148, https://doi.org/10.1192/bjo.2021.960. These examples establish objectless presentation but not exact phenomenological matching.

**4.** Modern philosophical formulations of the stronger hallucination argument appear in Frank Jackson, *Perception: A Representative Theory* (Cambridge: Cambridge University Press, 1977), and Robinson, *Perception* (1994). Intentionalism can grant common representational structure while denying that the structure becomes an inner object of awareness; the later chapters on transparency and the identity claim develop that response.

**5.** J. L. Austin, *Sense and Sensibilia* (1962), esp. secs. III and VII. Austin's central charge is not merely that the argument selects unrepresentative examples (though he presses that too — perception consists overwhelmingly of cases where things are as they appear), but that it depends on what he treats as a grammatical confusion: the move from "$x$ looks $F$" to "there is something that is

*F* of which the subject is aware." His priority claim is conceptual and linguistic rather than a claim about child development. For an extension of Austin's approach into epistemology, see John McDowell, "Criteria, Defeasibility, and Knowledge," *Proceedings of the British Academy* 68 (1982): 455–479. The decisive form of the diagnosis — recasting the confusion as a conflation of the *vehicle* of a representation with its *object* — is reserved for the later chapter that names that distinction.

**6.** Disjunctivism denies that veridical perception and subjectively indiscriminable hallucination share a common fundamental mental kind. J. M. Hinton introduced the disjunctive conception in "Visual Experiences," *Mind* 76 (1967), and *Experiences* (Oxford: Clarendon Press, 1973); Paul Snowdon, "Perception, Vision and Causation," *Proceedings of the Aristotelian Society* 81 (1980–81): 175–192, John McDowell, *Mind and World* (Cambridge, MA: Harvard University Press, 1994), and M. G. F. Martin developed importantly different versions. For the broader debate, see Alex Byrne and Heather Logue, eds., *Disjunctivism: Contemporary Readings* (Cambridge, MA: MIT Press, 2009).

**7.** Causally matching hallucinations supply the sharp pressure case. Martin does not simply deny a universal same-cause/same-effect law. He argues that if a matching hallucination receives a positive representational characterization sufficient to explain its salient features and effects, the same explanation threatens to screen off the naïve realist's world-involving account of perception. His response characterizes the matching hallucination negatively, through its indiscriminability from perception; see "The Limits of Self-Awareness," *Philosophical Studies* 120, nos. 1–3 (2004): 37–89, esp. 60–63 for the challenge and 71–73 for the response. Denying a common fundamental kind does not by itself deny shared psychological kinds; Martin's restrictive negative account concerns causally matching hallucinations, while other hallucinations may receive positive characterizations compatible with indiscriminability. The book favors an independently explanatory common representational character over that restricted negative account, not over a general denial of every positive mental feature in common.

**8.** The hallucinated clock adapts Bertrand Russell's stopped-clock example in *Human Knowledge: Its Scope and Limits* (London: George Allen & Unwin, 1948), 170–171, a predecessor of the justified-true-belief counterexamples Edmund Gettier sharpened in "Is Justified True Belief Knowledge?", *Analysis* 23 (1963): 121–123. Because the hallucinated belief remains causally uncoupled from the time, it fails influential safety and sensitivity conditions; see Ernest Sosa, *A Virtue Epistemology* (Oxford: Clarendon Press, 2007), Duncan Pritchard, *Epistemic Luck* (Oxford: Oxford University Press, 2005), and Robert Nozick, *Philosophical Explanations* (Cambridge, MA: Harvard University Press, 1981). Those diagnoses do not transfer automatically to the indirect realist's perceiver, whose inner states covary lawfully with the scene. The main text therefore uses the case only to isolate causal disconnection and then asks what, beyond lawful control, perceptual contact requires. It does not treat perception as a species of knowledge or convict indirect realism of Gettier luck.

## Chapter 4: The Transparency of Experience

> *"When we try to introspect the sensation of blue, all we can see is the blue: the other element is as if it were diaphanous."*
>
> — G. E. Moore, "The Refutation of Idealism"

---

CHAPTER OVERVIEW

Listen to a familiar song and try to turn from the sound to your hearing of it. Attention tends to slide back to the tune, just as a search for the seeing returns you to the wall, its color, shape, and place. I take this pattern, called the transparency of experience, as a problem for the claim that looking within reveals a private inner object. Blur, afterimages, and flashes behind closed eyes make the test harder, but they give our view work to do rather than hand the win to an inner screen. The result stays modest: looking within tells us how things seem, not yet what the seeing consists in. An inner-object theory may still explain those reports; it cannot simply claim the exercise reveals its truth.

### 4.1 The Window, Not the Screen

Listen to music for a moment—something you know well, something that moves you. Track the melody, feel the rhythm arrive, hear an instrument change the texture of the whole. The music presents itself as *out there*, coming from the speakers, organized in time as a thing in the world.

> Now try to shift your attention from the music to your *hearing* of it. Not the sounds, but the experience of sound. Not what you hear, but the inner event by which you hear it.

The attempt may return you to the music: a held note, the space between phrases, the way the sound fills the room. Hearing does not readily offer itself as a second sounding thing to inspect. Something billed as a screen behaves more like a window. If your report differs, keep it in view; the exercise asks what you find, not whether you can give the approved answer.

This observation has a name. Philosophers call it the *transparency* of experience, the phenomenon whereby perceptual experience, examined from the inside, ordinarily returns us to the objects of experience rather than to experience itself. The modern argument dates to 1990, when Gilbert Harman pressed the observation, in one short paper, into an argument with real edge: when we introspect a perceptual experience, we find only its *intentional objects* (the things the experience is directed at, the things it represents) and not the intrinsic features of the experience that make it represent those things.[1] The term sounds technical; the underlying observation stays stubbornly ordinary. The music made the point first; now take the standard visual case. Look at the wall in front of you and ask not what you see but what *properties of your experience* you notice. Attention ordinarily returns to the wall: its color, shape, and distance appear as features of the wall, not as paint on a second surface. You can also tell that you are seeing the wall rather than imagining

it. That is awareness of the mode in which the wall is presented, not acquaintance with another colored item between you and it. The experience does not vanish; it gives the world in a particular way.

### 4.2 What the Inner-Object View Predicts

The inner-object theorist has a reply ready: perhaps what you call "the wall" is an inner item presenting itself as external. The exercise alone cannot rule that out. On the sense-datum view, awareness immediately encounters a color-patch or sensory impression which in turn represents the wall.[2] The wall causes the episode; the datum supplies what awareness meets.

That view promises to explain why the same apparent scene can occur when the senses mislead us. The last chapter showed that error does not require an inner bearer, but it did not disprove every theory proposing one. Transparency tests a narrower claim: can the theory point to introspection as a direct disclosure of that bearer?

If it can, we ought to find more than the world merely seeming a certain way.

### 4.3 The Limits of Introspective Evidence

Suppose an inner-object theory claims not just that we encounter an inner item, but that introspection reveals it *as inner*. Then turning attention inward should disclose the bearer as private or mental. Redness, smoothness, or warmth merely appearing does not yet tell us what bears it.

Attention ordinarily returns to the wall, the music, the weight of the book in your hand. The colors and shapes appear as properties of a mind-independent wall, not as intrinsic properties of a mental bearer. This gives no independent support to the claim that introspection reveals the bearer as private. If a theory posits a bearer that appears as the wall, it owes an argument for that reading. The same report cannot serve as unmistakable evidence that its ontology has been directly inspected.

Harman put the point with his own example: attend to the intrinsic features of your visual experience of a tree, and the only features you find to attend to are features *of the tree* — its green, its height, out there, presented as belonging to it.[3] Look *through* the experience and you reach the world; look *at* it rather than through it, and no second colored bearer appears. The glass is clean. Nearly a century before Harman, G. E. Moore noticed the same pressure and gave it a fittingly optical name. Moore taught at Cambridge and would as soon defend common sense against a brilliant theory as agree with it. Trying to introspect a sensation of blue, he reported, one finds the blue and very little else; the sensation itself he judged *diaphanous*, see-through.[4] Moore immediately added that the conscious element could become distinguishable once one knew what to look for. He drew a different moral than the one this book will draw. What matters here is the first pull of attention: it goes to the blue. Harman gave that pull its modern dialectical edge.

Transparency reports what introspection finds; it does not yet tell us what experience consists in. Block's strongest reply posits no further object hidden from attention. His "mental paint" names a property of the experiential state whose nature, he argues, is not exhausted by what it represents.[5] The ordinary exercises provide no independent identification of such paint; they do

not rule out every proposed introspective route to it. Chapter 6 must supply the positive case that no constitutive residue remains.

Many readers hold a modern descendant of the picture: the brain builds a model of the world, transparent to its subject. As a claim about the *object of awareness*, the view faces a fork. If the subject gets the wall the model presents and never the modeling, the model belongs to the vehicle of seeing. If experience instead presents the model as another object besides the wall, transparency presses against it. Either way, calling the architecture a model does not settle Block's question about whether properties of the vehicle help constitute phenomenal character. Chapter 6 takes that up.

It would be cheating to run transparency only on the cases that flatter it. The real test comes from the awkward ones, and they have names.

1. Think of a page seen without the glasses needed to read it: the blur seems to be *in* the seeing, a property of the experience if anything is, since the page itself has not gone soft.
2. A colored afterimage can seem to drift across a wall, a patch of color no surface there wears.
3. Phosphenes, the flashes or swirls that can occur without light from a matching object, likewise seem to answer to nothing in the room. No need to produce them by pressing your eyes or staring at a bright source; their description gives us the problem.

If transparency holds, where do these live? They look exactly like the inner items the view denies.

Blur is the hardest case. You may see a sharp-edged page *blurrily* without seeing the page *as blurred*. The intentionalist must explain that difference by the determinacy, structure, or mode of a world-directed presentation without positing a fuzzy inner object. Afterimages present a related difficulty. An experience can present a colored patch at a region of the wall where no patch exists; in reflective cases, it can present the patch as an afterimage rather than as paint on the wall. The intentionalist owes an account of that apparent location which does not turn the patch into a mental object. Phosphenes raise the same demand in a starker form. On all these cases transparency must finally make good, against Block and against Boghossian and Velleman, who take blurry vision in particular to expose a residue of non-representational character.[6] Chapter 6 develops the positive thesis and tests the relevant distinctions against these cases; those arguments, rather than this first look, must bear the weight of the reply.

### 4.4 Vehicle and Content

That leaves the question the inner-object view was built to answer: what about illusions and hallucinations? The view's appeal drew heavily on cases where perception fails. If you're aware only of the world and not of inner objects, what are you aware of when the stick looks bent in water that isn't, or the building looks pink in a sunset, or, the limit case, when you hallucinate something that isn't there at all?

A full answer takes several chapters, but one distinction has to be visible before we can state the debt clearly.

> A neural process can serve as the *vehicle* of an experience. Its *content* specifies how things are presented: a stick as bent, a building as pink, a figure as standing before you.[7] The stick or building, when present, supplies the object; in hallucination no matching object need exist. The vehicle lies in the brain. This fact alone does not tell us where awareness lands.

Chapter 5 turns this distinction on the bent stick; Chapter 6 asks whether content can carry the weight placed on inner items.

### 4.5 The Exercise Is the Argument

The exercise has now crossed vision and hearing. When neuroscience says the brain "constructs a model of the world," the phrase invites an inner facsimile between perceiver and world. Transparency says: attend again; no separate screen ordinarily appears. The same caution will matter when we ask what a fluent machine sees. If the test demands introspective disclosure of an inner screen, humans have not passed it either. Whether we or machines have experience remains a different question.

I remember the night this first unsettled me — somewhere in the middle of working through Harman's paper for the first time, too much coffee and too little patience for philosophy that refused to be plain. The observation was almost embarrassingly simple, and it still took me a while to trust it, to resist the pull of three centuries insisting that something *had* to be there in the middle: some inner layer, some private what-it's-like-ness hovering between the mind and the world. Harman's observation felt faintly impolite, like pointing out that the emperor had been underdressed the whole time and no one had wanted to be the one to say so.

The exercise supplies the argument's negative half: no second surface announces itself as the bearer of what we hear or see. Even hallucination can present an apparent world without reaching an existing object. What experience therefore consists in remains to be earned. My dinner guests have learned to change the subject. You, I hope, will stay for another course.

### Chapter Summary

**The claim.** Perceptual experience has a transparent core. When we look within, sensory qualities usually appear as qualities of the music, wall, or tomato, not of a second inner bearer.

**Where the argument stands.** This first-person result recurs in ordinary cases, though people differ in their reports. It gives no independent warrant for identifying what appears as a private inner bearer. Blur, afterimages, and phosphenes leave the intentionalist with debts to pay, not victories already won.

**What's left open.** Transparency does not decide between every theory compatible with the report. Chapter 5 separates representing from an object represented; Chapter 6 must argue for what phenomenal character consists in.

**The hand-off.** The result leaves us with the difference between the *vehicle*, the state that represents, and its *content*, what that state presents. The next chapters use this tool to explain illusion

and hallucination without inner objects. For machine consciousness, it blocks only the demand that experience disclose a separate inner spectacle. It does not settle whether a human or machine has phenomenal experience.

**Footnotes**

**1.** Gilbert Harman, "The Intrinsic Quality of Experience," *Philosophical Perspectives* 4 (1990): 31–52. The paper is the locus classicus for the contemporary transparency argument. Harman's crucial move is to distinguish the properties *of* an experience from the properties the experience *represents its object as having*, and then to argue that introspection reveals intentional or relational features, not the intrinsic features responsible for the experience's content (38–39). This does not deny awareness of every property of an experience: Harman expressly allows awareness of its having a certain content. The observation itself has a longer pedigree — Moore's early work supplies the most direct precursor — but Harman gave it its modern dialectical role.

**2.** For a classic statement of the sense-datum view, see Bertrand Russell, *The Problems of Philosophy* (London: Williams and Norgate, 1912), ch. 1, "Appearance and Reality," especially the distinction between the sense-datum, immediately known in sensation, and the sensation as awareness of it. That distinction does not establish that introspection reveals its object as private or mental; the narrower challenge in Sections 4.2–4.3 applies only where such a disclosure is claimed. Tim Crane, "The Origins of Qualia," in *History of the Mind–Body Problem*, ed. Tim Crane and Sarah Patterson (London: Routledge, 2000), 169–194, sec. 2, reconstructs the further arguments through which immediate objects came to be treated as distinct from ordinary material objects. For the contemporary dialectic with naive realism, see James Genone, "Recent Work on Naive Realism," *American Philosophical Quarterly* 53, no. 1 (2016). Sydney Shoemaker, "Self-Knowledge and 'Inner Sense': Lecture III: The Phenomenal Character of Experience," *Philosophy and Phenomenological Research* 54, no. 2 (1994): 291–314, rejects an object-perceptual model while defending a qualia-compatible account; he is not an ally of the inner-object model.

**3.** Harman, "The Intrinsic Quality of Experience," 39. Harman's claim is that when Eloise attends to the intrinsic features of her visual experience of a tree, the only features she finds available to attention are features she experiences the tree and its surroundings as having — not features of the experience itself. For a sympathetic assessment of the argument's dialectical force, see Alex Byrne, "Intentionalism Defended," *Philosophical Review* 110 (2001): 199–240.

**4.** G. E. Moore, "The Refutation of Idealism," *Mind* 12 (1903): 433–453, at 450, on the "diaphanous" character of the sensation. Three cautions belong here. First, Moore deployed the observation in the service of a *realism about sensa* that this book rejects — the diaphanousness of the sensing was, for him, compatible with there being a sensum sensed; the transparency tradition after Harman turns the same observation against inner objects. Second, the word "diaphanous" has since been used in more than one way — Galen Strawson, for instance, uses it to mark something other than Moore's point (see "Real Direct Realism," 2015) — so the term should be handled with care. Third, and most relevant, Moore's very next sentence takes half of it back: "Yet it can be distinguished, if we look attentively enough, and if we know that there is something to look for." Moore himself judged the diaphanous element findable, and the anti-transparency literature (Kind and Stoljar, cited below) leans on precisely that concession. What the main text borrows from

Moore is therefore only the reported phenomenology, not his verdict on it: attend to the sensing of blue and you find the blue. For careful contemporary engagement, see Amy Kind, "What's So Transparent About Transparency?," *Philosophical Studies* 115 (2003): 225–244, and Daniel Stoljar, "The Argument from Diaphanousness," in *New Essays in the Philosophy of Language and Mind*, Canadian Journal of Philosophy Supplementary Volume 30 (2004): 341–390.

**5.** Ned Block, "Mental Paint and Mental Latex," in *Philosophical Issues* 7, ed. Enrique Villanueva (Atascadero, CA: Ridgeview, 1996), 19–49; and "Mental Paint," in *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, ed. Martin Hahn and Bjørn Ramberg (Cambridge, MA: MIT Press, 2003), 165–200. Block's mental paint is a property of the experiential state, not an object of an inner perceptual relation. It may represent redness without its nature being exhausted by that content; his "mental oil" names the stronger case of phenomenal features that represent nothing. Block grants the familiar tomato-attention report, distinguishes attention from awareness, and contests general transparency with other cases. The absence of a second object does not settle his property claim; Chapter 6 supplies the positive thesis needed for the reply.

**6.** The hard cases for transparency — blurry vision, afterimages, phosphenes — are developed later. On afterimages and other "phenomena without an object" handled representationally, see Michael Tye, *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind* (Cambridge, MA: MIT Press, 1995), ch. 3, and *Consciousness, Color, and Content* (Cambridge, MA: MIT Press, 2000), ch. 4, especially cases 3–5 in sec. 4.2. Paul Boghossian and J. David Velleman, "Colour as a Secondary Quality," *Mind* 98, no. 389 (1989): 81–103, and Ned Block, "Mental Paint" (cited above, n. 5), press blur against an exhaustive representational account. The reply must distinguish seeing a sharp object blurrily from seeing a blurred object clearly. Tye develops a determinacy-based treatment in *Consciousness, Color, and Content*, ch. 4, case 3. Crane's mode-based alternative in "Introspection, Intentionality, and the Transparency of Experience," sec. 4, instead rejects transparency understood as exhaustive object-and-content awareness; see note 7. These are distinct strategies for addressing the hard cases, not a settlement supplied by the ordinary exercise.

**7.** The vehicle/content distinction is drawn with particular clarity in Tim Crane, "Is There a Perceptual Relation?," in *Perceptual Experience*, ed. Tamar Szabó Gendler and John Hawthorne (Oxford: Oxford University Press, 2006), 126–146; see also Crane, "Introspection, Intentionality, and the Transparency of Experience," *Philosophical Topics* 28 (2000): 49–67, secs. 2, 4–5. In the latter, Crane defends strong intentionalism in a mode-plus-content form while explicitly rejecting transparency understood as the claim that introspection reveals only objects and their properties. Introspection can also reveal the mode in which the subject relates to content. That rejection does not contradict this chapter's weaker "transparent core," which allows mode-awareness and denies only that the ordinary exercise independently identifies a private bearer. Chapter 6 must defend the book's positive identity claim; Crane's distinction does not establish it.

# PART TWO — The World-Directed Mind

## Chapter 5: The Seeing and the Seen

> *"Look at a tree and try to turn your attention to intrinsic features of your visual experience. I predict you will find that the only features there to turn your attention to will be features of the presented tree, including relational features of the tree 'from here.'"*
>
> — Gilbert Harman, "The Intrinsic Quality of Experience"

---

CHAPTER OVERVIEW

Lay a map flat on the table: it can show a steep hill without making the table harder to climb. I use that difference to separate three things: the vehicle that does the work, the content that says how things stand, and the object it concerns, when one exists. Two mistakes blur those lines: we turn the state inside the skull into a thing we see, or demand a second bent thing to explain why a straight stick looks bent. Sight can get the public stick wrong without showing us a private replacement. I favor shared content across seeing and hallucinating because it offers one account of their apparent scenes, while the rival view gives the object a more basic role in seeing. That choice leaves a real debt: content must explain sensory presence without placing a screen between us and the world.

### 5.1 The Tool on the Table

Lay a walking map on the table and find the hill on your route. Crowded contour lines show a steep climb ahead, though the paper lies flat. You can smooth a crease in the map without making the walk any easier. Three things differ here: the ink and paper, how they depict the terrain, and the hill itself. The map can get the slope wrong; no second hill has to appear on the table to make that error possible.

The last chapter gave us transparency: looking for the seeing ordinarily returns our attention to what seems seen, rather than to a second colored surface.[1] The map helps name a distinction that observation needs. The physical **vehicle** does the representing; the **content** specifies how things are presented; an **object**, when there is one, supplies what the state concerns. We will use that distinction against two moves: treating the brain state as something seen, and demanding an inner bent thing because a straight stick looks bent.

The analogy has a limit. Looking at a map does not amount to seeing its hill. It shows only that a vehicle here can concern something elsewhere without sharing its properties. Direct perception needs more: the object must help cause the very presentation of it under the right world-involving conditions. The map earns the distinction, not a theory of sight.

### 5.2 The Grammar That Hides the Distinction

Open a dictionary to *experience* and you find, near the top, a harmless epistemic sense: *direct observation of or participation in events as a basis of knowledge.*[2] The definition posits no private object. Grammar nevertheless makes one remarkably easy to imagine. The events sit on one side, the observer on the other, and the noun offers us a third term: *experience*, ready to become the medium where the two meet.

The noun does not summon a ghost. It leaves a chair out for one.

Watch how the noun behaves. *Experience* takes adjectives, sits as a subject, gets *had* and *undergone* and *described*. Some adjectives properly qualify the episode: a recollection can be vivid, a headache painful, an afternoon rich. Trouble begins when the character of seeing gets pictured as an exhibit with its own colors and grain. There is the tree; there is Eloise looking at the tree; and now we imagine *Eloise's experience of the tree* as a third item that a sufficiently careful science might one day inspect.

Grammar can arrive with the stage already set. Once experience has been pictured as a third item, the famous problems nearly write themselves. How do the events out there get *translated* into properties in here? Why do the inner properties feel the way they do rather than some other way? How could anyone but the haver ever know about them? The noun did not cause three centuries of debate; it made the proposed scenery feel familiar before the arguments began.

None of this denies that experience has character. Looking at a sunset differs unmistakably from listening to a fugue: the redness, the late gold, the slow cooling of the light all belong in the account. The inner-exhibit picture prejudges where to look for that character. It sends us inward, to a third item whose own qualities supposedly make the difference. Chapter 4 found no such item announcing itself to attention. The positive explanation still lies ahead.

### 5.3 Vehicle, Content, and Object

Apply the map's distinction to seeing an apple on the counter.

> The **vehicle** is the physical state doing the work: here, a pattern of neural activity. The **content** specifies how the apple appears: red, round, and resting on the counter. These are its **accuracy conditions**, what would have to hold for the presentation to get things right. The **object** is the apple itself, when one exists and stands in the right relation to the seeing.

The content and object cannot trade places in this list. The same apple can look red or, under misleading light, brown; the object stays while its presentation changes. A hallucination may present a red apple where none exists. Content remains possible without a matching object, and neither case requires the neural vehicle to turn red.

The trouble starts when we let the vehicle's address determine where its object must be.

> Call the mistake the **container fallacy**: the vehicle of perception is a state of the brain, the brain is inside the skull, *therefore* what perception delivers must also be inside the skull: an inner item, not the outer world.

That *therefore* does not follow. The map lies on the table; the hill does not. Likewise, a neural state in your visual cortex can concern the apple on the counter without putting the apple in your cortex. Where representing happens does not by itself tell us where to find what it concerns.

Dress that fallacy in a modern lab coat and it fools serious people. *We do not really perceive the world,* the argument goes; *we perceive a representation in the brain, because the proximate cause of the experience is a brain state.* Fred Dretske caught the slip.[3] He had a working rule (explain the mind in terms continuous with the rest of nature) and a taste for the homely example that catches a grand mistake red-handed. A newspaper lies on your table and brings news of a flood elsewhere. You see its printed words; the flood does not thereby move onto the paper. The inscription's location and causal role do not fix where its subject matter lies. The analogy earns only that distinction: reading about a flood does not amount to seeing it. Likewise, a brain state's causal proximity does not by itself make that state what you see. Mistake the cause for the object and you have built yourself a screen.

### 5.4 The Argument from Illusion Comes Apart

Return to the straight stick in the glass of water.

Recall the argument that put you there, because its respectability carries the point.[4] A straight stick, half-immersed in a glass of water, looks bent. Nobody disputes that much; you can run it tonight at the sink. The argument continues: the stick itself remains straight. Pull it out and check; run your finger along it. And yet, the argument presses, *you are aware of something bent.* Bentness stands present to your mind. Since the public stick lacks it, some other object must bear the quality: an inner item, a sense-datum. Once that item gets its foot in the door, the continuous series from distorted to accurate perception spreads it across every case, and the world retreats behind a screen. That road runs from a stick in a glass to the veil of ideas, and for three centuries it has looked like a one-way street.

The argument turns on one step: from *the stick looks bent* to *you are aware of something that is bent.* It converts a presented quality into a bearer; the continuous series spreads the result. The container fallacy instead transfers the vehicle's location to what we see. One reifying family, two doors.

Of course bentness is presented. Sight presents the stick as bent and, in that respect, misrepresents the straight stick.[5] Bentness enters the accuracy conditions: the presentation would get that feature right if the stick had the shape it seems to have. Nothing yet requires the vehicle, or a second object, to bear the shape.

> Here the careful defender of inner objects digs in at his strongest. He grants the grammar. He grants that "looks bent" can smuggle. And he

presses anyway: *but I am genuinely presented with bentness; bentness stands there before my mind; presentation requires something that has it.*

This is the best version of the argument. A state can represent bentness with nothing bent anywhere, inner or outer, just as a sentence can concern a dragon without a dragon crouching behind the words. Representation needs no instance of its content to do its work.

The comparison leaves one debt unpaid. Reading about a dragon does not feel like having a dragon sensibly present, while the bend at the sink does feel given in vision. The defender can grant the point about sentences and press that difference. Fair enough. But the Phenomenal Principle now shows its hand: it no longer reports a neutral fact about what appears; it offers a theory of what sensory presence must involve. Until the bearer theory wins, the bent object does not follow.

Hunt on for an inner *thing*, itself bent, carrying the quality the way a sense-datum supposedly carries it, and you have made content into one more object seen. A straight stick, presented inaccurately; no second bent thing follows, in the glass or behind your eyes. And where on the smooth slope from distorted to accurate perception does awareness switch from an inner object to an outer one? Nowhere. The distorted end never earned an inner object in the first place.

Nobody saw the linguistic slide more clearly, or said it more enjoyably, than J. L. Austin. He delivered the lectures later reconstructed as *Sense and Sensibilia* at Oxford; G. J. Warnock assembled the book from Austin's notes after his death.[6] It takes the argument from illusion apart joint by joint, and two of its cuts help here.

1. Austin first exposes a grammatical sleight: the slide from "*x* looks bent" to "there is something that *is* bent which I sense." "Looks bent" no more requires a bent something than "looks likely" requires a likely something, or "looks like rain" requires a thing made of resemblance. The grammar of *looks* invites the inner object; it does not deliver it.

2. The second cut is one I had underestimated until I read him properly. The argument from illusion, Austin points out, runs together two things English keeps carefully apart: **illusion** and **delusion**. In an illusion, a real public arrangement looks other than it is: two equal lines look unequal, or a stage performer appears headless. In a delusion, by contrast, something wholly unreal may seem present: Austin's patient sees pink rats. The immersed stick belongs on the public-object side of that distinction, though Austin thought the case too familiar to count as a proper illusion without stretching the word. One real, straight stick looks bent through water. Nothing in that description conjures a private replacement.

This cut blocks the quick route from public misperception to a private object. A careful sense-datum theorist can still grant the stick throughout and demand a bearer for the presented bentness. That slower route returns us to the Phenomenal Principle, where the representational alternative has shifted the burden but not yet explained sensory presence. Austin clears the verbal undergrowth. He does not make the remaining argument disappear by taxonomy.

The argument from illusion therefore no longer forces the veil of ideas. "The stick looks bent" does not by itself place a bent object before awareness, and the continuous series cannot spread

an object that never entered at the distorted end. The inner bent thing required an additional theory of presence, quietly entered as a premise.

### 5.5 And the Theater Goes With It

The familiar theater makes the other mistake: it turns the vehicle into an object. Behind the eyes stands a private screen, with a Self in the dark watching the show. It can feel forced on us by sheer biology: *the brain sits in the dark, signals arrive, surely it must be working from an inner image.*

But a neural representation need not be an image anyone inspects. Neural states can register features, guide attention, and shape action without another viewing of those states. The container fallacy adds that viewing; the neural facts do not. Just as mapping the hill does not erect a hill on the table, processing its slope does not erect a slope inside the skull. We still need to explain how that processing lets us see. Adding a screen would give us another thing whose seeing needs explaining.

### 5.6 Disjunctivism, and Why the Book Declines It

Two camps agree that no veil stands between us and the world, and disagree about why. This book and the disjunctivists both send the inner object packing; they part over what perception shares across its good and bad cases. The sense-datum theorist and the representationalist both posit a positive **common factor** across that divide, then quarrel over what it involves — an inner object for one, representational content for the other. Which side is right decides whether the identity claim ahead has the materials it needs.

Disjunctivism plants its flag at the case Austin taught us to hold apart from illusion: the **hallucination**, the tomato that seems present when no tomato sits anywhere to be seen. It denies that genuine perception and a matching hallucination share one fundamental experiential nature.[7] M. G. F. Martin built its modern case. Really seeing the tomato amounts to a relation that *includes* the tomato, a state you could not occupy if the tomato were not there; the matching hallucination counts as a different kind of state altogether, sharing indiscriminability without sharing that basic kind. Either you stand related to a real object, or you occupy some other state that merely seems the same: a *disjunction*, with no neutral experiential core left for the sense-datum theorist to make inner. No screen, no veil; the relation reaches all the way out.

Disjunctivism makes for an elegant escape, and I respect it. It pushes realism about perception to its most uncompromising form, building the tomato itself into the constitution of the perception. This book takes another road, locating the world in what experience *presents* rather than in what *constitutes* it. Against the sense-datum theory we march together: two parties refusing to let the bent stick smuggle an inner object onto the scene.

Hallucination does not force either camp to admit an inner object. On the common-content account, a state can present a tomato where none exists; its accuracy conditions go unmet. No private tomato replaces the missing public one. That offers an alternative to the bearer theory, though sensory presence still needs explaining.

The disjunctivist has a complaint waiting in return. If one complete content-state can occur in both perception and matching hallucination, the tomato starts to look external to the experience's

fundamental nature — the veil rebuilt one level up. As stated the charge does not stick: content never becomes an object you perceive instead of the tomato. Whether a common content nevertheless leaves the current world too far outside the nature of seeing remains a second question, and the world-involving account ahead must answer it.

Why decline that road? A forged banknote helps locate the dispute. I may spend it because I take it for money. A reply in Martin's spirit says that its being indistinguishable from a genuine note explains that response; the two notes need not share a fundamental nature. Nor does a common belief about them prove a common kind of experience. The question concerns what best explains the apparent scene that prompts the belief, not just the belief afterward.[8]

I favor common content because it describes that scene positively: a tomato presented as red, round, and within reach, whether or not the tomato is there. It lets one account explain the detailed appearance while the right causal contact distinguishes seeing from hallucinating. Martin instead gives genuine perception the basic role and explains the bad case through its relation of indiscriminability to the good one. His account preserves the object's place within seeing more directly; ours promises a fuller common explanation of how things seem. That promise must earn its keep in the cases ahead. The banknote cannot decide between them.

So we let the disjunctivist go his own way, with a tip of the hat. He reaches the open air by another route. We keep the common factor and accept the bill: the positive theory now has to earn it.

### 5.7 The Seeing and the Seen

Right now you have something in view: this page, the wall behind it, the light in the room. The seeing does the presenting; its content specifies how the room appears. A vehicle works inside the head. The room need not move there, and content supplies no further exhibit standing between you and it.

Without an inner-object veil, the low-grade skeptical worry (*what if the screen and the world disagree?*) loses its grip. A second worry remains: whether common content leaves the current object outside the fundamental nature of seeing.

Naming the confusion blocks two arguments for an inner exhibit; it supplies no complete positive account. Seeing the apple's red still has a definite character. A serious view—I held a version longer than I like to admit—grants this chapter and insists on an intrinsic feel whose nature is not exhausted by what the state represents.[9] It does not rebuild the screen, and the next chapter owes it a careful answer.

If the felt character of your experience belongs to no inner exhibit, then what in the world does it consist in? Hold the question. Meanwhile, you can take the stick out of the water.

### Chapter Summary

**The claim.** One reifying family recurs across three centuries of trouble. The container fallacy turns an internal vehicle into an object of perception; the argument from illusion turns presented content into a further bearer. Both install a second object where representing needs none.

**Where the argument stands.** The vehicle's location does not place what we see inside the skull, and a presented bend need not belong to an inner bearer. Content specifies how things appear; an object, when present, supplies what the state concerns. Austin blocks the quick slide from public error to something conjured, leaving the Phenomenal Principle to defend its theory of sensory presence.

**What's left open.** What gives content sensory presence, how world-involving contact distinguishes seeing from hallucinating, and whether a common factor creates an interface all remain for the positive account. The qualia realist who still posits an intrinsic feel (Block's mental paint) gets a separate answer there as well.

**The hand-off.** Disjunctivism reaches the same anti-inner-object result by denying a common fundamental experiential nature. This book keeps common content for its explanatory work and must now show that content forms no new veil. The next chapter begins paying that bill.

**Footnotes**

**1.** The transparency observation this chapter inherits rather than re-runs belongs to Gilbert Harman, "The Intrinsic Quality of Experience," *Philosophical Perspectives* 4 (1990): 31–52, and, in its older form, to G. E. Moore, "The Refutation of Idealism," *Mind* 12 (1903): 433–453, at 450 (the "diaphanous" character of the sensing). The phenomenological demonstration — attend to the experience and you find only its object — is the business of the preceding chapter (*The Transparency of Experience*); this chapter assumes it and supplies the diagnosis.

**2.** *Merriam-Webster's Collegiate Dictionary*, 11th ed. (Springfield, MA: Merriam-Webster, 2003), s.v. "experience": "direct observation of or participation in events as a basis of knowledge." The definition itself carries no commitment to an inner object. Section 5.2 concerns the ease with which the noun can be reified, not a claim that lexicography created the metaphysics.

**3.** Dretske distinguishes representational vehicles from what they represent in *Knowledge and the Flow of Information* (Cambridge, MA: MIT Press, 1981) and *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995). The latter, ch. 2, distinguishes the object perceived from the fact learned through it: seeing a scale can disclose a fact about one's weight without making one's body the object seen. The newspaper example in the text is the book's illustration, not a quotation or an argument that reading and seeing share the same relation to their subject matter. It isolates the invalid inference from a vehicle's location or causal role to the location or identity of what it represents.

**4.** The argument from illusion is reconstructed at full strength, and assessed, in A. D. Smith, *The Problem of Perception* (Cambridge, MA: Harvard University Press, 2002), chs. 1–2 — the most thorough modern treatment, and notably one sympathetic to the argument's seriousness even where it ultimately resists the sense-datum conclusion. Smith isolates the principle the literature, following Howard Robinson, calls the "phenomenal principle" — that if one is sensibly aware of something as having a quality, then there is something of which one is aware that has that quality (Robinson, *Perception* [London: Routledge, 1994], 32) — as the argument's central premise. Intentionalists reject that inference: a state may represent a quality without anything's instantiating it. For the contemporary intentionalist diagnosis on which this chapter leans, see Tim Crane and

Craig French, "The Problem of Perception," *Stanford Encyclopedia of Philosophy* (Fall 2021 edition), and Crane, "Is There a Perceptual Relation?," in *Perceptual Experience*, ed. Tamar Szabó Gendler and John Hawthorne (Oxford: Oxford University Press, 2006), 126–146. The dragon case makes the alternative coherent; it does not by itself explain why perceptual content feels present. Chapter 6 takes up that debt.

**5.** This chapter locates the presented bentness in the *content* of a representation that misrepresents a real, straight stick. That move presupposes the same general representational structure across accurate and illusory seeing. Metaphysical disjunctivism need not contest that pairing; its characteristic denial concerns veridical perception and a subjectively matching hallucination. Section 5.6 explains why this book keeps a positive common factor across that stronger divide.

**6.** J. L. Austin, *Sense and Sensibilia*, reconstructed from the manuscript notes by G. J. Warnock (Oxford: Oxford University Press, 1962). The illusion/delusion distinction and the critique of the "looks *F*" ⟶ "is aware of something *F*" inference are developed in Lectures III–V. Austin separates illusion from delusion and objects to treating the familiar immersed stick as though it introduced a second bent object. That blocks the quickest route from misleading appearance to a private intermediary. A defender who explicitly adds a Phenomenal Principle or another bearer premise has made a further argument, however; Austin's taxonomy alone does not settle it.

**7.** The metaphysical (or "phenomenal") disjunctivism at issue in Section 5.6 is developed in M. G. F. Martin, "The Transparency of Experience," *Mind & Language* 17 (2002): 376–425, https://doi.org/10.1111/1468-0017.00205, and "The Limits of Self-Awareness," *Philosophical Studies* 120 (2004): 37–89, as well as "On Being Alienated," in *Perceptual Experience*, ed. Tamar Szabó Gendler and John Hawthorne (Oxford: Oxford University Press, 2006), 354–410; its modern origins lie with J. M. Hinton, "Visual Experiences," *Mind* 76 (1967): 217–227, and *Experiences: An Inquiry into Some Ambiguities* (Oxford: Clarendon Press, 1973), and with Paul Snowdon, "Perception, Vision and Causation," *Proceedings of the Aristotelian Society* 81 (1980–81): 175–192. The claim relevant here is not that veridical perception and subjectively indistinguishable hallucination share no true description or downstream effect, nor that indiscriminability entails positive phenomenal sameness. It denies a common fundamental experiential nature. Veridical perception receives a positive, object-involving account; Martin characterizes hallucination through its indiscriminability from genuine perception. A separate, *epistemological* disjunctivism, associated with John McDowell ("Criteria, Defeasibility, and Knowledge," *Proceedings of the British Academy* 68 [1982]: 455–479), concerns the reasons perception supplies rather than the metaphysics of phenomenal character. For the current taxonomy, including fundamental-kind, content, phenomenal, and epistemological formulations, see Matthew Soteriou, "The Disjunctive Theory of Perception," *Stanford Encyclopedia of Philosophy*, substantive revision July 28, 2026, secs. 2–5.

**8.** The common-factor pressure follows Tim Crane; see "Is There a Perceptual Relation?," in *Perceptual Experience*, ed. Tamar Szabó Gendler and John Hawthorne (Oxford: Oxford University Press, 2006), 126–146, and (with Craig French) "The Problem of Perception," *Stanford Encyclopedia of Philosophy* (Fall 2021 edition). The banknote contrast in the text extends that pressure. It distinguishes a neural event's efficient causal role from the subject's reason for belief and action, and both from whatever constitutes the experience itself. A disjunctivist may grant common

downstream beliefs or dispositions while denying a common experiential nature. The objection therefore asks whether that concession explains enough; it does not show that the disjunctivist has restored the disputed common factor under another name.

**9.** The deferral flagged in Section 5.7 concerns phenomenal character whose nature is not exhausted by representational content. Block's "mental paint" can itself represent redness; his stronger "mental oil" claim concerns phenomenal features that represent nothing. He explicitly distinguishes this stronger sense of "non-representational" from the weaker claim that representational content does not supply the whole nature of phenomenal character. Removing an inner *object* does not by itself answer either challenge about intrinsic *qualities*. The next chapter defends the positive identity claim and meets inversion and mental-paint objections; the hard-cases chapter then tests pain and mood, retaining unresolved debts concerning affective content and reasons for relief. These are arguments and applications to assess, not results established by the present diagnosis. Block, "Mental Paint," in *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, ed. Martin Hahn and Bjørn Ramberg (Cambridge, MA: MIT Press, 2003), 165–200. The paint/oil distinction was checked in the available manuscript transcription (dated January 27, 2004, manuscript pp. 8–9); the corresponding location in the cited volume has not been independently verified.

## Chapter 6: The Identity Claim

CHAPTER OVERVIEW

What makes seeing red feel unlike seeing blue? I defend a strong answer: the feel of a conscious state consists in the full way it shows us a world. That profile includes what the state presents, whether it works through sight or bodily feeling, how much detail it gives, how its scene is laid out, and what draws notice. The claim says what the feel consists in; which states are conscious at all is a separate question with its own tests. Blur, afterimages, shifts of attention, and cases of reversed color give us no need for a private coat of mental paint; the tomato, not an inner copy, remains what we see. Match two conscious states in this full, well-defined profile and let them feel unlike, and my claim fails.

The last chapter ended on a question: if the felt character of experience belongs to no inner object, what does it belong to?

An identity claim says that two descriptions have picked out one thing. Lightning is the useful case, because you can know the flash and not know what it is. The flash and the electrical discharge are not two events, one causing the other. The flash is not made out of the discharge the way a statue is made out of bronze. One event, reached by two routes. You discover the identity. The meanings of the words do not settle it. Water and $H_2O$ have the same form. The clear stuff in the glass and the chemical compound are one substance under two descriptions. Nothing in the word *water* tells you the compound.

That is the form. It is not yet a claim about experience.

Write the identity as an equation, with the feel on one side and something physical on the other, and the question becomes where the equals sign lands. This book rejects a separate mind-stuff pushing the body by a route the rest of nature does not record, but refusing that picture still leaves views that keep the sign off anything physical. Property dualism retains phenomenal properties joined to the physical by irreducible laws. Illusionism disputes what our judgments of feel tell us. Keith Frankish preserves sensory discrimination, aversion, attention, and introspective reports, and denies that the states we report possess genuinely phenomenal properties; his *quasi-phenomenal* properties explain why we take them to do so. I take a different position: some judgments of how experience feels correctly recognize a physical, world-presenting character. The quarrel with Frankish concerns whether those judgments are true, not whether people wince or tell red from green. I mean to defend that identification, not win it for lack of rivals.[1a]

For the theories that do write the equation, four places are enough to see the choice. The sign can land on the cell, the firing that accompanies the ache. It can land on the role, the job a state does, whether or not any tissue of ours does that job. It can land on a nature hidden inside matter, something the equations of physics name only by what it does. Or it can land between the feel and what a conscious state presents. That last place is the one I will defend. Within a conscious state, phenomenal character consists in its complete world-presenting profile: what it presents and how it presents it, one and the same thing under two descriptions, nothing over and above it.

The name for that place is strong representationalism, and the cases that follow are what it has to survive.

Two familiar pictures are missing from the list. The inner theater makes the immediate object of awareness a private copy, and the last chapter closed it. Qualia realism locates the feel in an intrinsic quality that representation cannot capture; it remains open, and Section 6.2 takes it up. The work of this chapter is to show why the feel consists in what a conscious state presents, rather than in the cell, the role, or a nature hidden inside matter.

### 6.1 The Older Observation: Moore, Brentano, and the Diaphanous Mind

*Experience* easily becomes a third thing between observer and world, yet attention to that supposed exhibit slides through it and lands on the world. Harman found as much when he sent Eloise to introspect her seeing of a tree and she came back with nothing but the tree.[12]

The observation is old. G. E. Moore, in 1903, called the experience-relation *diaphanous* — see-through — noting that when we try to introspect a sensation of blue, "the other element is as if it were diaphanous."[3] Moore did not think the act invisible, though: he added at once that it "can be distinguished if we look attentively enough." Aquinas had treated the form in the mind as what cognition works *through*, and Brentano made that outward directedness the mark of the mental. Neither they nor Moore held the identity claim defended here; the family resemblance lies in where they looked to specify the act. What an experience is *like* turns on what it is *of*. So this chapter asks what *experience* actually picks out, once the noun stops pretending to name a thing over and above the world it gives us.

### 6.2 One Observation, Two Claims

Attend to your visual experience of a tomato and you expect to find some inner item: a mental image, a private color-patch, a shimmer of redness behind the eyes. Chapter 4 ran that exercise. Nothing turns up but the tomato.[4]

The observation bears on two distinct claims. The inner-theater picture puts a private representation between perceiver and tomato: a screen behind the eyes, watched by a Self, from which the world gets inferred. Qualia realism need not do that. It can treat phenomenal character as an intrinsic property *of the perceptual state* whose nature content does not exhaust — Block's mental paint — while granting that the tomato remains the object of awareness.[5][6]

Demolishing the theater therefore leaves the identity question open. Strong representationalism says phenomenal character consists in the complete intentional profile and nothing further; if mental paint does independent explanatory work, the claim fails. Section 6.3 gives the positive case, the inversion cases test it, and Chapter 7 follows the proposed residue into pain and mood.

### 6.3 The Positive Case: From Transparency to Content

The case has three stages. Each needs more than the stage before it supplies.

1. **Transparency undermines the supposed inner exhibit.** Introspection returns the tomato, wall, music, or headache as presented; it does not disclose a private object or coat of phenomenal paint.

2. **The wider evidence supports representational determination.** Changes in perception and attention can be measured against a fixed stimulus and compared with reports of changed appearance. These comparisons support an account of the phenomenal contrasts through the world-presenting profile, though no experiment measures that whole profile without theoretical assumptions.
3. **Identity and historical width require further choices.** The profile might ground character rather than be one and the same thing. Or character might consist in its present structure, while history fixes a wider reference. I adopt a wide identity.

Start with transparency. Michael Tye has spent thirty years refining the thought behind it: introspection provides our primary access to experience from within, yet systematically directs attention to what experience presents. If an inner object or intrinsic paint figured in experience's basic structure, some access to it would be expected. Its persistent absence counts against both pictures and supports a world-directed account.[7]

Tye separates the next move himself. Transparency first creates trouble for qualia realism. He then adds a *further plausible hypothesis*: the qualities experience presents fix or ground its phenomenal character. Even that hypothesis initially yields only dependence — what he calls weak property representationalism, a supervenience claim. Strong property representationalism makes the additional identity claim. In plain terms, "the feel never changes without the presented qualities changing" still differs from "the feel consists in those qualities." The chapter must earn the stronger conclusion rather than slide into it.

Nor does transparency decide among every physicalist theory. Higher-order and self-representational accounts explain consciousness by adding a representation of the first-order state, so they too enter the comparative contest.[7b] The first-person evidence needs discipline. Dennett's *heterophenomenology* takes subjects' reports and discriminations seriously while refusing to treat their metaphysical interpretation as incorrigible.[7c] A report that the tomato appears red constrains the theory. It does not inventory the state's ultimate constituents.

The second stage asks which account best explains the wider pattern. Take a state already counted as conscious and borne by an identified subject. Its phenomenal character, the book proposes, consists in its complete world-presenting intentional profile: independently fixed content under a perceptual or bodily-affective mode, with determinacy, field structure, and attentional organization.[8] This profile exceeds a bare object-label. It records whether the tomato appears red or green, clear or blurred, centered or peripheral, seen or imagined, commanding attention or fading behind the mug. Content-fixing history, the test for state consciousness, and the identification of the subject remain separate questions.

Mode names the way a state presents its content, not another property attributed to the tomato. Sight and imagery can present the same shape through different causal routes, with different roles in inspection and correction. Those differences help specify mode without defining it by the feel. Attention enters the profile where its organization changes the presentation, not merely which answer gets selected. Neither term licenses an extra intrinsic quality. Each needs the independent constraints the tests below require.

Transparency does not prove that every intrinsic feature has been found wanting; a hidden feature remains logically possible. Kind and Stoljar press exactly there. Kind distinguishes *strong* transparency, on which we cannot attend to experience except by attending to what it presents, from *weak* transparency, on which such direct attention is difficult but not impossible. On her reading, the transparency argument for representationalism needs the strong version. Stoljar likewise argues that diaphanousness underdetermines representationalism. A qualia realist can grant the tomato exercise and retain a property of the state that ordinary introspection overlooks.[9] Blur, double vision, and afterimages can even seem to offer indirect access to that property. Block adds cases of his own: a near finger can look doubled while the eyes focus on a distant object, closed eyes in daylight still offer a visual field, and a jackhammer may enter awareness before it draws notice.

None of Block's cases requires an inner object. Double vision, on this account, changes how many objects seem present or where they seem to lie. That proposal needs the same independent checks as blur; naming a change in content does not establish one. Kind shows that we can attend to *how* the world gets presented. She does not thereby isolate a property outside the presenting. Whether the presenting profile exhausts each case remains the further question, and the tomato exercise cannot answer it alone. So the intentionalist owes more than a new description of each apparent exception. Content must earn accuracy conditions through tracking, producer-consumer organization, and world-involving history; mode, determinacy, field structure, and attention need constraints not supplied by the phenomenal report alone. Otherwise "the complete profile" would mean only whatever made the theory come out true.

The strongest tests run in both directions. Preserve state consciousness, subject, and the independently specified profile while phenomenal character stably changes: that breaks representational determination. Change the complete profile while character demonstrably stays fixed: that challenges the stronger identity, since determination alone can allow different profiles to yield the same character. The tests can guide experiments. Stimulus control, discrimination, generalization, error patterns, and intervention can constrain the profile. Recurrence, flexible reuse, correction, and cross-system influence can constrain state consciousness. Perturbation can help identify the bearer. First-person judgment remains evidence rather than an infallible reading of inner furniture.

The question is how to specify that profile without quietly reading the answer off the feel. Blur gives us a place to start: the printed edge stays put while the eye's grip on it changes. Before asking how the page looks, we can ask which differences in its lettering the visual system can still resolve. The worked comparison below separates those measurements from the phenomenal report.

### 6.3.1 The Paint's Home Turf: Blur, Afterimage, Attention, Imagery

Tye reports Block's challenge through a cinema screen: seeing a blurred picture clearly can differ from seeing a sharp picture through unfocused eyes. Can a theory of content capture the difference?[10] The everyday version needs only a page and reading glasses. Without the correction, nearby lettering can go soft while you know perfectly well that the print has not changed. You reach for your glasses rather than for a sharper book.

That knowledge does not settle what your visual state presents. You can know the Müller-Lyer lines have equal lengths while they keep looking unequal. So the intentionalist cannot infer "no visual error" merely from your correct belief about the print. The proposal to test is more specific: in this case, blur consists in a loss of determinacy about the contours, rather than an added smear or a positive presentation of fuzzy ink.

Consider a controlled comparison, not a report of a completed experiment. Keep the same viewer, page, distance, lighting, and fixation point; vary the optical correction. First establish, with correction in place, which small shifts of a letter's edge the viewer can distinguish. Then repeat with the altered optics, varying edge position and spacing across trials rather than asking whether anything *feels* blurred. Include trials where the ink itself grades gradually into the paper, so a genuine diffuse boundary has its own measurable profile. Randomize positions and responses to check that guessing or a preference for one answer does not supply the effect.

This gives "less determinate" some content before the phenomenal verdict arrives. Suppose the viewer can still locate the letter but cannot distinguish nearby placements of its edges. The hypothesis under test reads the state as making a coarser spatial claim. A state making precise but scattered estimates could earn the same overall score, so error patterns across trials, and what the viewer does next, must help tell the two apart. Its proposed target remains the printed contour: moving that contour changes the responses, and restoring focus restores the finer discrimination. The relevant error would put the contour outside the region presented. Silence about its exact position need introduce no error. Visual mode remains fixed by the ongoing optical input and its role in guiding inspection, not by calling the result "visual." Fixation constrains the field, while performance on the same location task helps constrain attention. None of this proves that every aspect of attention stayed fixed, or that discrimination exhausts experience. It prevents those terms from doing whatever we please.

Now ask how the page looks. Suppose reports of softness track the independently measured loss of edge detail and both reverse when focus returns. The report contributes the phenomenal side of the comparison; it did not define the loss we measured. The intentionalist has explained a contrast through the visual system's changed presentation of the page. But if a diffuse-ink condition and an altered-optics condition match on these measures while looking different, that is a problem to investigate, not permission to announce an unmeasured difference in "mode." The complete-profile claim would survive only if some further difference earned independent support.

Block's physicalist can accept the whole pattern. The changed optics alter a neural feature whose intrinsic character constitutes the blur; its representational effects explain the diminished discrimination. No second screen or spectator enters that reply. A still leaner nonidentity rival can grant the intentionalist's entire profile and say the feel obtains *in virtue of* it. Neither rival must predict a mismatch, so the comparison cannot single out identity. What it does supply is a constrained explanation of the contrast, and it weakens the original claim that content has nothing to say about blur.

You may reasonably think I have explained what changes without yet explaining why it feels different. My answer identifies the altered presentation with the altered feel; it does not promise

a further event after the edges lose their precision. That answer earns a preference when the same account explains the accuracy, errors, discriminations, and phenomenal contrast without a separately specified qualitative mechanism. Necessary grounding can match that success. Its proposed direction of dependence, rather than an imaginary canvas, supplies the remaining disagreement taken up in Section 6.5.3.

Afterimages and phosphenes bring the same question into cases without a matching colored object. An afterimage can appear to drift across a wall; a phosphene presents a spark without a light source there. If you take the patch for something on the wall, the account is straightforward: the visual state attributes a color at a location where nothing instantiates it. The representing, not some private patch, is illusory.

The harder case begins once you know perfectly well that you have an afterimage. Boghossian and Velleman put the point exactly: the patch appears *there* without appearing actually to exist there.[11] The identity claim should not answer by pretending you have made the simple mistake you just disavowed. Tye's better reply distinguishes levels. Your nonconceptual visual state still presents something red and filmy before the background; your higher-level judgment represents that very presentation as unreal. My comparison is with the Müller-Lyer lines, which can keep looking unequal after you have measured them. Recognition corrects the judgment without repainting the lower-level seeing. Redness and apparent location therefore remain in the visual content, while the subject's assessment correctly denies any red thing there. No visual-field object has been discovered — only a fallible presentation and a judgment that catches it failing. Remove the object and you do not strand the paint; you get a representation with empty hands, plus a perceiver who knows they are empty. The patch itself will fade before the argument does.

Attention supplies an actual experimental comparison. Marisa Carrasco and her colleagues studied the question with two patches of tilted stripes, briefly shown on opposite sides of a fixation point.[10a] A preceding dot drew attention to one location without predicting either patch's contrast or tilt. Observers reported the tilt of whichever patch looked higher in contrast. Cueing changed which physical contrasts they judged equal; within part of the tested range, it also improved their accuracy in judging tilt. The experiment thus separated a measurable stimulus property, an intervention on attention, an appearance judgment, and performance with a right or wrong answer.

The controls matter. Lengthening the interval after the cue removed the appearance effect. Asking which patch looked *lower* in contrast reversed the choice, rather than leaving observers mechanically choosing the cued patch. These results count against a simple preference for whatever received the cue. They do not settle every possible decision-bias account. Better tilt discrimination constrains processing without by itself proving that contrast looked stronger; the appearance judgment supplies evidence of that further change. If attention really makes unchanged contrast look higher, the account must accept a small misrepresentation. If attention changes only the decision, this experiment supplies no phenomenal contrast for identity to explain. Either way, "attention changed" cannot serve as a blank check on content.

The imagined tomato differs along another axis. The image comes gappy, unstable, thin where the percept runs dense (try counting the seeds in your imagined tomato). It arrives under an imaginative mode, summoned and dismissed at pleasure. Perception answers to the world whether you like it or not. Both states can influence belief, but under different warrants and through different causal routes. Same object, perhaps; nowhere near the same profile. Across blur, afterimage, attention, and imagery, the proposed paint has yet to explain a stable phenomenal contrast the independently specified profile misses.

That is the case for representational determination, made on the ground where the paint should be strongest. Three questions remain for later. First, explanatory exhaustion still falls short of identity. A residue-free critic may grant every result above and say that the profile *grounds* or *constitutes* phenomenal character. That position adds no paint, and transparency gives no reason to reject it. Section 6.5.3 therefore makes the third-stage argument openly: absent an independently motivated causal, modal, or persistence difference, identity carries one property where grounding distinguishes profile from character and adds an asymmetric dependence relation. Second, Chalmers's pressure remains. Transparency closes neither the epistemic gap nor the modal gap; conceiving a profile without its character may seem to separate them even when no actual case does. Chapter 9 asks what conceivability establishes, and Chapter 11 asks why two concepts of one fact can resist an *a priori* bridge. Third, these cases vary present processing or mode. None holds the complete current structure fixed while testing a change in historical property reference, so they support the wider intentionalist program, not a preference for historical width over a structural account. Inverted Earth and the historyless duplicate test that further claim. I retain the wide verdicts, but with less warrant than the broader intentionalist account, and their revisionary consequences cannot return as evidence for the premise that produced them. The positive result stands. Economy favors identity for now, and I rank it as the best explanation of the actual cases without claiming that introspection has settled either of Chalmers's questions.

## 6.4 Three Objections

A family of classical objections presses on the identity claim, each naming a feature of experience that content supposedly cannot capture. Three belong here — the cases that grant the experience a worldly object and dispute only whether its content fixes its felt character. I take them in order, meeting each here and sending the fuller apparatus to the back matter where a case earns it.

### 6.4.1 Mary's Room

Frank Jackson's Mary's Room, the most famous objection of the family, now has a chapter of its own.[12] Here I record only how the identity claim bears on it. Mary knows every physical fact about color vision from inside a black-and-white room, then walks out and sees a tomato. She learns something — but what she gains is a new *way of representing* a property she already knew, the perceptual mode of presentation that one acquires only by undergoing the state.[13] Whether that concession costs physicalism anything is Chapter 8′s question, and Chapter 8 answers it with more care than a summary can.[14]

### 6.4.2 The Inverted Spectrum

The family's oldest member comes next: the inverted spectrum. Chapter 2 filed a folk version of this case among the puzzles the theater generates — a private inner display, swapped behind the scenes, unstatable once the theater falls. The version that survives the demolition needs no inner items; it states its challenge in the identity claim's own vocabulary. Suppose your color experience runs systematically inverted relative to mine (you see red where I see green, and the reverse), yet we were both raised to call the tomato "red," so nothing in what we say or do betrays the difference. If that is coherent, phenomenal character cannot consist entirely in representational content, since our experiences would represent the same properties while feeling different. You have probably asked it yourself, as a child if not since (*do you see my red?*), and the pull deserves respect: the difference, if real, seems locked away where no test could reach it. The reply starts with the color solid, the three-dimensional arrangement of colors by hue, brightness, and saturation. Its lopsided geometry makes a full swap that preserves every relation harder to sustain for human vision. That puts pressure on the example; it does not prove inversion impossible. Block can ask about creatures engineered without those asymmetries. Their case must stand on whether the stipulated match is possible, not on the geometry of our color vision.[14a]

### 6.4.3 Inverted Earth

The inverted spectrum imagined two people inside one world. Inverted Earth moves one of them to another world, and it is the harder case. The scenario comes from Block, who built it as the version representationalism cannot absorb.[15]

> Picture it. One night, without your consent, two things are done to you at once. Fine color-inverting lenses are slipped behind your eyes, and you are carried to Inverted Earth, a planet exactly like ours except in its colors. The sky there reflects what we would call yellow. Ripe tomatoes throw back the reflectance of grass. The locals, raised on an inverted vocabulary, call that sky "blue" and those tomatoes "red." The two inversions cancel. You wake and notice nothing wrong. The sky overhead looks the way the sky has always looked to you, you call it "blue" as your new neighbors do, and nothing inside you has had to adapt. Let years pass. Your color words now follow local usage. Add a stronger condition: learning and consumer use have recruited your visual states to track the local reflectances. This changes their perceptual property content, not just your vocabulary. Stipulate that the complete present structure of discrimination, attention, and memory comparison matches the earlier one. You remember the childhood sky and notice no difference between that blue and this one.

This is a theoretical construction, not a combination established in a laboratory. Long residence and successful action do not by themselves establish the perceptual recruitment we have granted. Given that stronger condition, the wide content shifts while the subject notices nothing. The

inverted-spectrum reply cannot help: here the present color-space relations remain fixed. The question concerns the changed worldly address, not a demonstrated change of feel.

The disputed claim concerns sameness of character across the two times. The unchanged report supports it, but does not settle it. Nor does a change in wide content settle the opposite verdict. Two questions come apart.

Does the case restore mental paint? No. A structural representationalist can grant Block the report and still have no inner colored item. She can hold that the complete synchronic profile — color-space relations, discriminations, expectations, attention, memory dispositions, and rational control — fixes how the sky looks, while causal history fixes the state's wider reference. On that account the present organization continues to present the same structural color even after the wide address changes. Both that view and the book's phenomenal externalism reject an inner colored item, and both send attention outward. They disagree over whether historically fixed reference belongs to phenomenal character. Transparency cannot decide between them.

If character stays fixed while its represented color changes, wide identity fails. The structural rival offers a reason to expect that result: every present relation that shapes discrimination, attention, and comparison remains fixed. My wide account offers another explanation. The very color presented also fixes what would make the presentation accurate. Section 6.5 gives the boundary that this claim needs: property attribution enters character; singular object reference and the referent of a kind-concept do not.

Both theories therefore draw a boundary within content. The wide view unites appearance with the property that makes it correct or mistaken. The structural view explains why matched present organization should preserve appearance despite a change in wider reference. Blur and attention support both. I have no independent comparison here that establishes the wide view's advantage when only history differs. I retain it as a less-supported extension of the broader intentionalist case, not as the winner of an experiment neither side has faced.

On that view, memory keeps its old object: the image remains *of* the earthly sky. The color it presents that sky as having re-anchors with perception. Character changes without detection because the standards of comparison move with it. This remains the book's Inverted Earth verdict. Its defense shows how the verdict can cohere with the unchanged report; it does not confirm the predicted change. Section 11.3 draws the consequence for acquaintance.

Two further objections press here: bodily sensations like pain, and moods that seem to feel like something while being about nothing. Chapter 7 takes up both. A bodily address helps explain phantom pain, but pain's badness and the content of mood remain unfinished applications. The perceptual cases support the identity claim more securely than those affective extensions do.

### 6.5 What the Thesis Commits Us To

Strong representationalism does not say consciousness amounts to nothing or that felt life comes as an illusion. It says the felt dimension consists in the complete world-presenting profile of a conscious state. The redness you see remains real: your visual system presents the tomato's

surface as having a property it has, or, in error, one it lacks. No private coat of color need accompany that presentation.

It also commits the book to **intentionalist direct realism**. When you see the tomato, awareness reaches the tomato immediately; no representation stands in its place. Content interposes no object; it states how the world stands, and in successful perception the current object satisfies it through the right causal-historical connection. Searle supplies one influential route to that compatibility, making the perceived object part of the experience's causally self-reflexive conditions of satisfaction. The book takes the compatibility lesson without taking over his precise machinery.[16]

That commitment needs more than the slogan that content is transparent. Three things must stay apart. By **vehicle**, I mean the neural episode that does the representing, produced through the causal chain from light at the retina to activity along the visual pathways. The **content under a perceptual mode** specifies how that state presents the world and what would make the presentation accurate or mistaken. Successful seeing takes the tomato itself as its **intentional object**, the thing the experience presents; a pure hallucination has no current mind-independent object in that role. These three enter one account, but they do not change places. Neither vehicle nor content becomes something seen merely because the theory needs it to explain seeing.[17]

Hallucination tests the distinction. Suppose a hallucination and a perception share the same presented content and perceptual mode. One conclusion follows: content alone does not suffice for seeing, since it can occur unsatisfied. A second conclusion does not follow: that when a tomato stands before you, you first perceive the content and then infer the tomato from it. In the good case, the tomato causes and satisfies the presentation under the right world-involving relation; it is what the experience is of. In the bad case, the same presentation runs unsatisfied. Shared content explains the matched character and rational pull. Satisfaction and current coupling explain why only one episode counts as seeing. No private item enters as the object of awareness or as a premise from which the subject reasons outward.

Relationalists press the hardest version of this. Bill Brewer and M. G. F. Martin both argue that the mind-independent object partly constitutes the experience, so that nothing content-like could have occurred unchanged without it.[18] Charles Travis presses from another side, arguing that experience issues no verdict determinate enough for the world to satisfy, and John Campbell that acquaintance with an object must come first, before any content about it. The reply is that common content explains the matched character of perception and hallucination, while satisfaction and current coupling explain why only one of them counts as seeing.

Perceptual content can remain nonconceptual and more or less determinate; the theory does not make experience whisper a sentence. Its accuracy conditions come from independently measurable work. Producer-consumer organization tracks something; learning corrects it; some variations alter discrimination and control; and one current object, not another, stands in the right world-involving relation to the state. A subject need not formulate those conditions any more than a cat needs the concept *trajectory* to misjudge a jump. The causal-historical coupling fixes the token object; it does not borrow a demonstrative concept from the perceiver. Chapter 15 owes the full account. If matched history and control leave no stable correction pattern that distinguishes

the candidate contents, concede Travis's silence at that grain. If demonstrative reference survives systematic breaks in current object-coupling, Campbell's priority claim gains ground.

Content has two layers here. Causal coupling makes this tomato, rather than its twin, the object seen. It settles which thing the experience gets right or wrong. The property layer presents a red, round, glossy surface at a place and distance. Replace the tomato with an exact twin while preserving that presentation and the look stays the same. A matching hallucination preserves the property layer without any tomato to fill the object role. I locate character in this layer, under its mode of presentation, not in the identity of the object.

The boundary concerns the work a content does, not whether its source lies inside the skull. A property attribution specifies how the scene is presented: red here, round there, an edge at this distance. A singular reference specifies which thing fills that role. A kind-concept classifies what we encounter through a further practice of recognition and judgment. Seeing the glass *as water* may change attention, grouping, or salience. On this account it changes character only through such changes in the property layer. The concept's referent adds no further feel. PANIC's *abstract* condition in Section 6.5.1 marks the exclusion of particular objects, not the exclusion of all history.

Hilary Putnam built Twin Earth to make a point about words. Its lakes contain XYZ, a liquid with water's familiar appearance but a different chemistry. Oscar and Twin Oscar live in 1750, before either knows that chemistry. Their words *water* reach different kinds, and their sightings reach different lakes. Yet the lakes look alike. Our account predicts that agreement because the two states present the same perceptible properties under the same mode. It does not infer that every fact about character lies inside the head.

Why count historical reference to red while excluding reference to $H_2O$? The proposed test starts with the perceptual system's own work. Consider sorting colored tiles under changing light. Test which changes in the surfaces alter sorting and which changes in the light the system discounts. Compare novel tiles, corrected mistakes, and the history through which the sorting was learned. These facts can support a property attribution that applies across different objects. They do not require the sorter to know the word *red* or the physics of reflectance. Vision can track an optical property without presenting its scientific description.

Now exchange water for XYZ while holding every such perceptual relation fixed. The bare change in chemistry supplies no new visual contrast or correction rule. A community can still fix its word's reference through samples and deference. That route explains a difference in concepts, not a difference in what vision presents. Mere causal contact with $H_2O$ does not make its hidden chemistry part of the look. If learning changes what the viewer picks out or attends to, we must test that change in the property layer instead.

This test constrains what counts as perceptual content. It does not show that the property's historical reference, rather than the current pattern of contrasts, fixes character. The structural rival can accept the tile example and keep that reference outside character. Expert recognition remains another challenge: the book must explain any changed look through the property layer, rather than grant a kind-concept its own feel. Twin Earth therefore leaves the wide and structural accounts tied. Inverted Earth and Swampman separate their predictions, but do not independently

establish which prediction is right. Chapter 16 returns to what Twin Earth does show about words.[18a]

Block's physicalist paint proposes a different bargain. An intrinsic property of the neural state supplies the feel, while content explains accuracy and rational direction. No second screen follows. For example, the same change in visual processing can explain both lost edge detail and the page's blurred look. The intentionalist identifies the look with the altered presentation; a vehicle theorist identifies it with the neural feature that realizes that presentation. Our blur comparison does not separate those explanations. A rival need not predict a missed contrast merely to compete. Such a contrast would favor it, but an independently supported account of the vehicle could also do explanatory work.

I prefer the representational identity because it connects the character of a state with what that state presents as true of the world. Take a white tile seen as red under odd light. The red attribution tells us both how the tile looks and what the view gets wrong. A structural account can also explain the look, then use the state's wider relations to assess its accuracy. My account uses the represented property for both jobs. That unity motivates the choice; it does not establish that two explanations track two properties. Nor does it show that the structural division costs more. Where history varies but all present structure stays fixed, the evidence considered here leaves that comparison open.

The book pairs this thesis with a separately defended realism about color. The tomato is red, and its redness belongs to it — a *type of surface spectral reflectance*, genuinely there to be reached. Strong representationalism alone does not tell us which properties colors are. Error theory, dispositionalism, primitivism, and perceptual variation all challenge the reflectance account.[18b] On the combination defended here, the felt redness consists in presenting the surface as carrying a reflectance-type it really carries. One world, color included.

One debt comes due here. The directness just defended runs entirely through *aboutness*, and this chapter leaves that aboutness a primitive, exactly where its ally Searle leaves it. Chapter 15 develops a naturalistic answer: selection, learning, tracking, producer-consumer organization, and continuing world-involving use can make a state answerable to what it concerns, whether the mechanism evolved or was engineered. The state can misrepresent because the function or use that fixes its standard can fail. This supplies a general model with worked cases, not a completed account of every pain and mood.[19]

### 6.5.1 The Consciousness Profile

The phrase *of the right kind* now has to earn its keep. Brains carry many representations that nobody experiences. Early vision registers edges and motion; priming alters a later judgment; bodily regulation tracks temperature, glucose, and pressure. Sophisticated influence on behavior cannot by itself mark consciousness, since unconscious processing can discriminate, predict, and control. If every representation counted as felt, the identity claim would spread consciousness wherever information flowed. That would erase the distinction the theory was introduced to explain.

The phrase conceals four questions unless they are separated.

1. **Content fixing**: what does the state present, and what makes that presentation accurate or mistaken? World-involving selection, learning, tracking, and producer-consumer history belong here. Chapters 12–16 take up that debt.
2. **Character determination**: among conscious states, what makes this one look red rather than green, blurry rather than sharp, painful rather than neutral? Content under a perceptual or bodily-affective mode, with its determinacy, field structure, and attentional organization, supplies the proposed profile.
3. **State consciousness**: why does this content figure in experience rather than remain unconscious processing? The house proposal is a distributed access-and-control profile, specified below. Influence on behavior alone does not satisfy it.
4. **Subject individuation**: for which persisting system is the content poised and integrated? A subject cannot enter as a proprietor added to the machinery. It must be identified through the organization across which perception, memory, expectation, belief, desire, planning, correction, and action hang together. Chapter 17 develops that organization; Chapter 24 tests the criterion where candidate systems are nested or distributed.

The identity relation adds no fifth ingredient. It says that, within a content-bearing state already counted as conscious and borne by an identified subject, phenomenal character consists in the complete world-presenting intentional profile — nothing over and above it, and no further paint waits to be attached. The remaining empirical questions ask which architectures fix content, which states meet the consciousness profile, and which persisting control organization bears them.

Failures at one level need not spread to the others. A conscious state outside the proposed distributed profile would refute that account of state consciousness while leaving open what fixes its character. A mistaken subject boundary would force a better account of the bearer, not new mental paint. Conversely, a matched-profile phenomenal contrast would strike the identity claim even if the book had correctly identified both conscious states and both subjects. Keeping the questions apart prevents one success from being spent twice.

Here is the positive proposal. A content-bearing state is conscious for a subject when it participates, *as that content*, in a recurrently sustained and selectively available pattern of influence across the subject's integrated economy. Under suitable conditions it can guide attention, working memory, inference, planning, report where report is possible, and flexible action; feedback from those activities can preserve, revise, or displace it. No one output is necessary. Breadth, flexible reuse, recurrence, and two-way correction do the discriminating work together. A subliminal prime that bends one response, or a regulator that adjusts one variable, lacks the profile even though it affects behavior.

The proposal replaces a doorway with a causal pattern. A state can acquire greater or lesser reach, stability, and control; borderline cases should therefore be expected. Global broadcast, local recurrence, higher-order monitoring, and predictive organization become hypotheses about how physical systems realize parts of the profile, not rival definitions among which the book refuses to choose. Dennett's Multiple Drafts model supplies the right architectural lesson: contents undergo distributed revision and some gain what he later called "fame in the brain" by influencing memory,

report, reasoning, and action, with no final presentation to a witness.[19a] His denial of intrinsic qualia does not amount to a denial that pain or experience occurs. The book shares that refusal of mental paint while defending its own identity claim about the complete world-presenting profile.

Tye's 1995 framework, **PANIC** — *Poised, Abstract, Nonconceptual, Intentional Content* — gives conscious content four conditions.[20]

1. *Poised* content stands available for use in forming beliefs or desires.
2. *Abstract* content concerns property types rather than particular tokens.
3. *Nonconceptual* content does not require its subject to possess concepts for the features it presents. The shades we can tell apart far outnumber the color concepts we command.
4. *Intentional* content reaches toward properties of the world or body.

The last three conditions specify the form and target of perceptual presentation. Poise specifies its place in the subject's economy. It requires no current attention, verbal report, or actual judgment. A headache can shape the day before its sufferer names it. A crowded scene can look richer than the few items a subject manages to report. Poise instead names the standing availability at the center of the distributed profile. PANIC helped isolate that feature; it does not by itself fix the content's worldly address, identify the persisting subject, or specify the recurrence and cross-system correction that make availability more than a latent disposition.

That profile rules out two easy counterfeits. A thermoregulatory mechanism can register the coffee's heat and adjust a hand without presenting the heat within a wider cognitive economy. A conceptual thought can enter reasoning while lacking the fine-grained perceptual mode of feeling warmth. A subject, here, is not whatever already has experience. It is the relatively maximal persisting control organization within which content remains available across memory, expectation, belief, desire, correction, planning, and action. Competing goals get settled there. Altering the content alters later system-level control. Chapter 17 develops that nonphenomenal test and Chapter 24 applies it to nested and distributed systems.

Two pressures test that proposal: Ned Block's overflow argument and the live contest among global-workspace, recurrent-processing, and higher-order accounts over how the profile gets realized. Block argues that phenomenal detail can exceed cognitive access, not merely what a subject actually reports. In partial-report tasks, a later cue can make any one of several rows available without making all rows available together. That leaves room for selective availability, but it does not show that access constitutes consciousness. Nor does it settle whether inaccessible detail was already experienced. Those are empirical disputes the house proposal must face. Find a stable conscious state that lacks the distributed profile, or an unconscious state that matches it in content-sensitive breadth, recurrence, flexible control, and correction, and the state-consciousness account fails even if the identity claim survives.

### 6.5.2 Where an Identity Can Drop

The opening named four places and left the names for here. The cases are now in. They have favored an identity with the presented profile. They have not closed every other drop. The refusals have not changed. Substance dualism, property dualism, illusionism, and behaviorism still drop no sign. What follows says what the cases can pressure.[21c]

Type identity gets the physicalist demand right. One world, and the feel is not a second substance. Smart wanted the identification strict: the kind, pain, and a kind of brain process. Place had already called consciousness a brain process. He meant the is of composition, a cloud and its droplets, not a definition. That composition is still an identity, and still a claim about the kind. What it identifies is the feel with the vehicle. The contrasts in this chapter moved with what was presented: content, mode, determinacy, field structure, attention. That tracking does not show the vehicle is not the feel. Water does not look like $H_2O$, and a neuron need not look like the ache for Smart's identity to be the one he wanted. That theory still needs an independently supported account of which neural property supplies the feel. It need not predict a contrast the profile misses merely to remain a rival. The cases here leave the carrier and the content varying together. The vehicle is also not what the experience presents. That is transparency. It is not yet the identity.

Anomalous monism gets the particular right. In 1970 Davidson identifies this pain with this physical event, and denies any strict law joining the kinds. The particular event can be physical. The kind of feel is still unnamed. A token identity without a pair for the character does not say what the ache consists in.

Realizer identity gets the octopus right. The physical occupant of a causal role may differ by species, so an octopus may hurt without our C-fibers. The openness to other tissue should survive. It does not name the ache. The occupant is still a vehicle of the role. Identifying character with that occupant leaves the same question: which physical feature realizes the feel, and what supports that claim?

Machine functionalism gets a different lesson, and a real one. The mental state need not be a kind of hardware. Putnam identified it with the machine-table role. A causal diagram can be drawn without yet being a program. Computational functionalism adds that running the right program is already the state. Neither a role nor a program, just as such, fixes what a state presents. Causal roles help specify the mode in our account; they do not supply its worldly content on their own. Chapter 20 keeps the openness to other tissue and rejects both as a settlement of what a state concerns. This chapter's evidence already requires a presentation. A diagram does not supply one.

Russellian identity and powerful-quality identity get a lesson the structural vocabulary of physics really does leave open. They ask after the concrete nature of the property that has the powers, and they drop the sign inside matter. Russellian identity identifies the categorical basis of a physical role with a phenomenal property, rather than building a conscious whole out of smaller ones. Powerful-quality identity identifies a property's qualitative nature with its causal power. Applied to a conscious state, that would place the felt character in the powers of the physical property. The contrasts tracked here did not require that drop. They moved with the presentation, and a presentation can be carried by a physical property whose nature the structural vocabulary never names. Not requiring the drop is not the same as closing it. Chapter 10 meets the full case. Constitutive panpsychism stays off the table. Constitution would leave two relata and the gap between them.

Reductive and non-reductive physicalism sort some of these rows. They are not further places to drop the sign. A reductive physicalist identity reduces a mental type to a physical kind, or to

the physical occupant of a role. Smart and the realizer row do that. Davidson refuses the type reduction and is the non-reductive case. Machine functionalism reduces a type to a role. That role is not a physical kind, and the reduction is not Davidson's refusal. Supervenience without identity drops no sign. Searle's formula, caused by and realized in, is that refusal, and Chapter 11 says why it cannot close the gap.

The comparison narrows the field. It does not force the sign off the physical item that carries the content. A role or a program can be shared by a system that presents nothing, and Davidson's token still leaves the kind of feel unnamed. The vehicle that realizes a presentation varied with that presentation in the cases examined here, so the cases do not separate the carrier from what it carries. Strong representationalism names the pair the contrasts tracked. A state's phenomenal character consists in its complete world-presenting profile, including its manner of presentation: one and the same thing under two descriptions, nothing over and above it. The name is the field's. It does not hand the row to any one author. It does not reduce the feel to C-fibers, and it does not delete the feel. Part Three still has to show how a physical state acquires that content. Tracking the profile is not yet that identity. The next section says why.

| View | What it identifies | Where the sign drops |
|---|---|---|
| Type identity | A sensation-type and a brain-process type. The textbook example pairs pain with C-fiber firing. | The neural kind. |
| Anomalous monism | A particular mental event and a particular physical event. No strict law joins the kinds. | The token, not the kind. |
| Realizer identity | A mental state and the physical occupant of its causal role. The occupant may differ by species. | The occupant. |
| Machine functionalism | A mental state and the abstract causal role itself, the machine-table state, not the hardware. | The role. |
| Computational functionalism | A mental state and a computational state. Running the right program is the state. | The program state. |
| Russellian identity | On the identity reading, the categorical basis of a physical role and a phenomenal property. Not a whole built from its parts, and not thoughts in particles. | The intrinsic nature of matter. |
| Powerful-quality identity | A property's qualitative nature and its causal power. One property, two descriptions. | Inside the physical property. |
| Strong representationalism | A state's phenomenal character and its complete world-presenting profile: content under its mode, with determinacy, field structure, and attention. The character consists in that profile: one and the same thing, nothing over and above it. | Between character and profile. |

The table records the pairs. It does not replace the meetings in later chapters, and it is not a survey of every identity in the field. It is the set this chapter can pressure. A higher-order theory that

identified the feel with the thought about the first-order state would drop the sign in a place this table does not name. The higher-order theory the book meets does not say that. It says the extra thought is what makes a first-order state conscious, which is the question of Section 6.5.1. Integrated information theory identifies experience with a cause-effect structure. A sensorimotor identity would identify seeing with the exercise of a skill. The enactive account in view has the skill constitute the experience. Place's composition, kept in the identity family above, identifies the sensation with the process: one thing, described by its makeup. Noë's verb leaves the skill and the experience as two relata unless a further identity is stated. The word does not sort them. The number of relata does. Phenomenal intentionality uses the last row's pair but disagrees about what fixes the content. Since "consists in" means identity here, reversing the order would change nothing. The difference lies elsewhere: on that view phenomenal character is the source of intentional content; on this book's view world-involving history fixes content independently of character, which keeps the identity from being circular. None of those claims was tested by the contrasts above. Leaving them off the table is not an answer.[21d]

### 6.5.3 The Identity Adopted

On the cumulative case, the book adopts the identity claim and the distributed profile as defeasible accounts of phenomenal character and state consciousness. The comparison just made narrows the rivals. It does not close the last step. Take a conscious state borne by an identified subject. Its phenomenal character consists in the state's complete world-presenting profile: independently fixed content under a perceptual or bodily-affective mode, with determinacy, field structure, and attentional organization. The profile takes in the properties presented, not the particular object that satisfies them; that object settles accuracy (Section 6.5). The character and the profile make one and the same thing, reached by two routes; the claim does not set a feel atop a profile that supports it. Content-fixing history supplies the address; distributed access and control mark the domain; integrated persistence identifies the bearer. These claims constrain the identity without becoming extra ingredients in phenomenal character.[21]

The claim goes beyond correlation, supervenience, grounding, or constitution. Like the identity of water with $H_2O$, it would hold necessarily if true while remaining knowable only *a posteriori*. The analogy gives the logical form, not evidence that this instance must succeed.[21a]

The strongest remaining rival grants the entire explanatory result without adding mental paint: the profile explains every phenomenal contrast, but *grounds* or *constitutes* phenomenal character rather than exhausting it.[21b] Parsimony against inner objects cannot answer that view. The empirical evidence underdetermines the last step.

The grounding rival can make a stronger argument than "it still seems different." The optical and cognitive organization explains why the page looks blurred, she says; the blur does not explain why that organization obtains. Her theory preserves that explanatory direction while granting necessary covariation and adding no independent mental paint. Grounding need not come with different possible-world or persistence conditions. Its purpose is precisely to distinguish what depends on what even where both must occur together.

The book's reply separates causal explanation from metaphysical priority. Changed optics can explain the onset of one state, and a detailed account of its operation can explain a contrast, even if the state has two descriptions. The causal direction from removing the glasses to altered presentation does not establish a second direction, from that presentation to a distinct phenomenal fact. Nor does our ability to understand one description first establish that what it describes comes first in nature. Grounding could earn that further direction through an independently supported account of what phenomenal character requires. The blur and attention evidence has not supplied one.

The preference for identity therefore concerns explanatory commitments, not a census of invisible objects. Both accounts carry the same demanding empirical profile. They do not carry the same necessity. If the feel and the presentation make one thing, the feel cannot vary without the presentation, because nothing varies apart from itself. If they make two things, something must tie them together in every possible case, and grounding names that tie without explaining it. Grounding also asks us to accept a direction of dependence that the present comparisons do not distinguish from an order of explanation. What identity leaves unexplained lies in us: why our two ways of thinking meet only after inquiry. Chapter 11 takes that up. I prefer to identify the feel with the presentation that already accounts for its contrasts. A matched-profile phenomenal difference would defeat representational determination; an established modal separation would defeat identity; an independently motivated grounding asymmetry would remove this economy advantage. The book chooses the account that adds no unexplained tie, while leaving the eraser on the desk.

That economy argument compares identity with grounding. It does not choose wide identity over a structural identity, which adds no second felt property. Return to the tile: both can explain why it looks red. Both can use wider relations to say what the view gets wrong. They differ over whether the wide reference helps make the look itself. I take it to do so. Blur and attention have not tested that choice. The evidence is stronger for an account through what perception presents than for the historical width I adopt. Inverted Earth drift and Swampman's initial lack of experience inherit that weaker support. Using the identity to deflate the hard problem cannot increase it.

### 6.6 The Thesis and the Hard Problem

Transparency closes neither the epistemic gap nor the modal gap. Chalmers's question — why should physical activity feel like anything at all? — survives the disappearance of an inner screen, and Chapter 11 meets it directly.[22] The present chapter has defended a conditional claim: phenomenal character consists in the representational profile of the right kind. It has not made that identity obvious from the armchair or ruled out every conceivable separation.

If the identity holds, however, asking why the profile *additionally* produces a feel double-counts one event. The remaining difficulty concerns our concepts and access: neural organization described from outside and lived experience undergone from within disclose the same process without yielding an *a priori* bridge. The necessity therefore carries a real intellectual cost. It does not require a second property or law in nature.

A reader who grants only that the profile grounds the feel still keeps much of this. No second substance enters, and no mental paint. If the grounding holds necessarily, a zombie remains impossible. The hard problem then shrinks rather than dissolves: one question remains, why this profile grounds this feel. The identity claim removes that question; grounding leaves it standing.

The older picture makes that cost look like a production problem. Physical processing occupies one column, private experience another, and the philosopher asks how the first could generate the second. The identity claim refuses the second entry: once the relevant physical activity takes the form of poised, world-directed presentation, phenomenal character consists in that activity's complete profile. What remains hard concerns why nature contains such organized activity, how brains realize it, where its conscious range begins, and why our phenomenal and physical concepts meet so reluctantly. None of those questions disappears. They no longer require an ontological bridge between two properties already counted separately.

The same separation disciplines the machine question. We must ask what a system's states present, whether any meet the distributed state-consciousness profile, which persisting organization bears them, and what phenomenal character consists in for any conscious state thereby identified. Nothing in those questions demands carbon or a mammalian silhouette.

Part Four inherits the dependency rather than escaping it. Its conclusions about content and understanding rest on history, use, and integration; its consciousness corollary rests here. A stable phenomenal difference between fully matched conscious profiles would defeat the identity claim and that corollary together. Programhood supplies no substitute argument.

**Chapter Summary**

**The claim.** Once a state counts as conscious and an identified subject bears it, phenomenal character consists in its complete world-presenting profile — one and the same thing, not a feel resting on a profile beneath it. This profile includes independently fixed content under a perceptual or bodily-affective mode, with determinacy, field structure, and attention. No inner paint remains beyond it.

**Where the argument stands.** The opening asks where the sign drops, in four plain places. Section 6.5.2 names the pairs. The cases pressure a role or a program that need not present anything. They do not separate the carrier from the content. Blur and attention show how stimulus, processing, and appearance can be compared without defining every variable by the feel. They support the representational account but do not choose uniquely between identity and necessary grounding; the preference for identity adds a defeasible judgment about explanation. The same evidence also leaves wide and structural identity tied. Inverted Earth and Swampman separate their predictions, not their observed success. I retain the wide verdicts with less support than the broader intentionalist account; rejecting them need restore no mental paint.

**What's left open.** Causal history and producer-consumer organization can fix content where a world-involving history privileges a correctness condition, though its exact grain may remain underdetermined. The distributed access-and-control profile marks state consciousness; competing theories propose its physical realization. Integrated persistence identifies the subject, and only then does identity specify character. A change in feel between matched conscious profiles would

defeat the identity; a matched conscious/unconscious pair would defeat the state-consciousness account. Borderlines and bearer boundaries remain questions for science rather than hidden additions to content.

**The hand-off.** Pain and mood seem to offer mental paint its safest refuge. The next chapter asks what pain and mood present, distinguishing what the account explains from what it still owes.

**Footnotes**

**1a.** Keith Frankish, "Illusionism as a Theory of Consciousness," *Journal of Consciousness Studies* 23, nos. 11–12 (2016): 11–39, secs. 1.2 and 1.6–1.7. Strong illusionism denies phenomenal properties, including purported physical phenomenal properties, rather than merely rejecting intrinsic private qualia. It retains quasi-phenomenal properties and permits a functional sense of what experience is like. The house view identifies phenomenal character with the independently specified world-presenting profile; it does not add an ineffable quality to the organization Frankish preserves. Appendix B.16 locates the dispute in whether restricted phenomenal judgments can truthfully concern that profile and states where the contrast could become verbal.

**1.** *Merriam-Webster's Collegiate Dictionary*, 11th ed. (Springfield, MA: Merriam-Webster, 2003), s.v. "experience" — the same entry the preceding chapter quotes. The grammar I describe (*experience* as a noun that takes adjectives and gets *had* and *undergone*) runs across the entry's senses and across the corresponding entry in the *Oxford English Dictionary* (3rd edition).

**2.** Gilbert Harman, "The Intrinsic Quality of Experience," *Philosophical Perspectives* 4 (1990): 31–52, esp. 39. The Eloise-and-the-tree example sits at the heart of Harman's transparency argument; the paper is short and worth reading whole, especially for its move from the introspective datum to the negative thesis about intrinsic non-representational properties of experience.

**3.** G. E. Moore, "The Refutation of Idealism," *Mind* 12 (1903): 433–453. Moore develops the transparency observation at 446 and at 450 calls the awareness element of the sensation, not the sensation as a whole, "diaphanous," adding at once that it "can be distinguished if we look attentively enough, and if we know that there is something to look for." Moore himself therefore rejected the strong thesis that introspection finds nothing but the object; the chapter borrows his observation, not his verdict on it. Brentano's thesis that intentional inexistence marks the mental is at Franz Brentano, *Psychology from an Empirical Standpoint*, trans. Antos C. Rancurello, D. B. Terrell, and Linda L. McAlister (London: Routledge, 1995 [1874]), Book II, ch. 1, p. 88. The Aquinas comparison is deliberately limited: *Summa theologiae* I, q. 85, a. 2 treats the intelligible species as that *by which* the intellect understands, with the external thing as its primary object; see Robert Pasnau, *Theories of Cognition in the Later Middle Ages* (Cambridge: Cambridge University Press, 1997). That is a family resemblance about cognitive direction, not an anticipation of representational identity. For contemporary scrutiny of transparency, see Amy Kind, "What's So Transparent About Transparency?," *Philosophical Studies* 115 (2003): 225–244, and Daniel Stoljar, "The Argument from Diaphanousness," in *New Essays in the Philosophy of Language and Mind*, ed. Maite Ezcurdia, Robert J. Stainton, and Christopher Viger, *Canadian Journal of Philosophy*, Supplementary Volume 30 (Calgary: University of Calgary Press, 2004), 341–390.

**4.** The transparency datum is canonically introduced in Harman, "The Intrinsic Quality of Experience"; refined in Michael Tye, *Ten Problems of Consciousness* (Cambridge, MA: MIT Press, 1995); developed further in Alex Byrne, "Intentionalism Defended," *Philosophical Review* 110 (2001): 199–240. For a recent overview, see Tim Crane and Craig French, "The Problem of Perception," *Stanford Encyclopedia of Philosophy*, Fall 2021 edition.

**5.** Ned Block, "Mental Paint," in *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, ed. Martin Hahn and Bjørn Ramberg (Cambridge, MA: MIT Press, 2003), 165–200, esp. 170–73. Block's view differs from classical sense-data theories: mental paint counts as a property of mental *states*, not as the object of an inner perceptual relation. The distinction matters for Section 6.2′s argument that qualia realism and the inner-theater picture are separable targets.

**6.** Block, "Mental Paint," sections 1–3, makes the distinction between qualia-as-properties-of-states and qualia-as-perceived-inner-objects explicit. His paint consists of phenomenal properties that represent without having their nature exhausted by representational content; he reserves *mental oil* for phenomenal properties that do not represent. A sophisticated qualia realist can therefore reject the inner theater without conceding the identity claim. Block's opening discussion also challenges transparency through the unattended jackhammer, daylight through closed eyelids, and the appearance of a near finger while attention remains on a distant object. These cases challenge the inference from outward attention to phenomenal exhaustion, not merely the idea of a private picture. For the position the book ultimately rejects but acknowledges as the strongest qualia-realist option, see also Galen Strawson, "Real Materialism," in *Real Materialism and Other Essays* (Oxford: Clarendon Press, 2008), 19–52.

**7.** Michael Tye, *Consciousness, Color, and Content* (Cambridge, MA: MIT Press, 2000), ch. 3; and Tye, "Transparency, Qualia Realism and Representationalism," *Philosophical Studies* 170, no. 1 (2014): 39–57, https://doi.org/10.1007/s11098-013-0177-8. The later essay makes the sequence explicit. Transparency creates trouble for qualia realism; the claim that represented qualities fix or ground phenomenal character enters as a "further plausible hypothesis." Tye distinguishes weak property representationalism, a supervenience thesis, from strong property representationalism, an identity claim. The main text preserves those steps rather than treating transparency as a shortcut to identity.

**7b.** Higher-order and self-representational theories offer physicalist alternatives to the first-order account of state consciousness. Higher-order theories require a suitable representation of oneself as being in a first-order state: David Rosenthal, *Consciousness and Mind* (Oxford: Clarendon Press, 2005); Hakwan Lau and David Rosenthal, "Empirical Support for Higher-Order Theories of Conscious Awareness," *Trends in Cognitive Sciences* 15, no. 8 (2011): 365–373, esp. 365–367, doi:10.1016/j.tics.2011.05.009. The higher-order representation need not itself be conscious, so this requirement introduces neither a spectator nor an automatic regress. Uriah Kriegel's self-representationalism instead places the world-directed and self-directed aspects within one conscious state: *Subjective Consciousness: A Self-Representational Theory* (Oxford: Clarendon Press, 2009). The substantive dispute concerns whether such self-representation is necessary and whether changing it can change consciousness while first-order content remains fixed. Lau and Rosenthal explicitly distinguish subjective awareness from objective task performance; matching

one discrimination score does not establish matching the complete first-order organization. Misrepresentation supplies a further test: when the higher-order state characterizes its target incorrectly, or lacks a target, the theory must specify which representation determines conscious character. See Ned Block, "The Higher Order Approach to Consciousness Is Defunct," *Analysis* 71, no. 3 (2011): 419–431, and Rosenthal's reply in the same volume. Following the higher-order representation offers a substantive answer rather than a contradiction, but it requires an account of why its content fixes appearance while the first-order state's content governs the relevant discrimination. The house proposal declines a necessary higher-order condition pending evidence that this additional explanatory role cannot be supplied by the distributed access-and-control profile. Its comparative claim remains defeasible; transparency alone does not decide between these theories.

**7c.** Daniel C. Dennett develops heterophenomenology in *Consciousness Explained* (Boston: Little, Brown, 1991), ch. 4, esp. 72–98. The method treats a subject's sincere reports as data from which to construct the world as it appears to that subject, while withholding assent from the subject's metaphysical interpretation of what the reports disclose. The present chapter adopts that neutrality about inference without importing Dennett's entire account. His rejection of qualia targets purported additional intrinsic, private, ineffable properties; ch. 12 and the Appendix preserve content-bearing discriminative states and the reality of pain. The book's own positive thesis identifies phenomenal character with the independently specified world-presenting profile, while treating reports about it as fallible evidence within the same causal order as the state reported.

**8.** The thesis remains identity rather than correlation or mere covariance, but the inference to it is abductive and defeasible. The identified relatum for phenomenal-character determination is the independently specified world-presenting intentional profile — content under a perceptual or bodily-affective mode, with determinacy, field structure, and attentional organization. Three neighboring questions remain distinct. World-involving producer-consumer history fixes the content and its accuracy conditions; Chapter 15 develops that account. The distributed access-and-control profile determines which content-bearing states fall within the identity claim's domain; global-workspace, recurrent-processing, higher-order, and other proposals bear on its realization. Integration and persistence identify the system that bears the state; Chapter 17 develops that account. None of these distinctions adds a proprietor, and none permits the phenomenal profile to expand after the fact merely to absorb a difference. The evidential claims must also be separated: the ordinary perceptual comparisons constrain representational determination, while neither they nor transparency select historical property reference over a complete synchronic structural account of character. The adopted wide identity therefore exceeds the best-supported portion of the argument.

**9.** The objection that the transparency datum *underdetermines* the representationalist conclusion — that one can grant the phenomenology and deny the thesis — is pressed by Amy Kind, "What's So Transparent About Transparency?," *Philosophical Studies* 115 (2003): 225–244, esp. 230–33. Kind distinguishes weak from strong transparency and also presses questions about transparency's scope; the main text isolates the strength distinction relevant to the identity argument. See also

Daniel Stoljar, "The Argument from Diaphanousness," in *New Essays in the Philosophy of Language and Mind*, ed. Maite Ezcurdia, Robert J. Stainton, and Christopher Viger, *Canadian Journal of Philosophy*, Supplementary Volume 30 (Calgary: University of Calgary Press, 2004), 341–390. The representational reply to this underdetermination objection is developed by Michael Tye, "Transparency, Qualia Realism and Representationalism," *Philosophical Studies* 170, no. 1 (2014): 39–57, esp. secs. 6–7. The book adds a constrained parsimony claim: once content has been fixed without reading it off the feel, a residue compatible with every report yet explanatory of none has no work left to do.

**10.** Michael Tye, *Consciousness, Color, and Content* (Cambridge, MA: MIT Press, 2000), ch. 4, "Case 3: Blurred Vision," 79–83, reports Block's then-unpublished cinema-screen example at 80 and answers it through representational indeterminacy. The argument distinguishes representing an object's boundaries as indefinite from failing to specify their precise location. At 82–83 Tye allows phenomenal indistinguishability in some suitably matched cases and develops a watercolor comparison, credited to Frank Jackson, for cases with a phenomenal difference. Block's general defense of phenomenal properties not exhausted by representational content appears in "Mental Paint," in *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, ed. Martin Hahn and Bjørn Ramberg (Cambridge, MA: MIT Press, 2003), 165–200; the cinema example here is attributed through Tye, not located in that essay. The edge-discrimination comparison in the main text is the book's proposed test of the determinacy account, not an experiment reported by either author. Discrimination constrains a candidate content attribution; it neither exhausts the conscious profile nor establishes phenomenal sameness. Tim Crane, "Introspection, Intentionality, and the Transparency of Experience," *Philosophical Topics* 28, no. 2 (2000): 49–67, sec. 4, offers a different intentionalist treatment: blurry content can remain while mode distinguishes its presentation from the subject's taking the world to be that way. His mode-based reply should not be identified with Tye's indeterminacy reply.

**11.** Paul Boghossian and David Velleman, "Colour as a Secondary Quality," *Mind* 98, no. 389 (1989): 81–103, use the recognized afterimage to press a projectivist difficulty: the colored patch appears at a location without appearing to be a colored object located there. Michael Tye answers in *Consciousness, Color, and Content* (Cambridge, MA: MIT Press, 2000), ch. 4, by distinguishing the nonconceptual visual content, which remains illusory, from the subject's higher-level conceptual assessment that no such object is present. Tye compares this conflict with a hallucinated pink elephant recognized as unreal. The Müller-Lyer comparison in the main text is the book's gloss on that distinction. The main text follows the two-level reply rather than treating every recognized afterimage as a flat judgment that a colored object occupies the wall.

**10a.** Marisa Carrasco, Sam Ling, and Sarah Read, "Attention Alters Appearance," *Nature Neuroscience* 7 (2004): 308–313, esp. 309–312, doi:10.1038/nn1194, reports attention-induced changes in apparent contrast. The uninformative peripheral cue, contingent orientation task, and points of subjective equality are described at 309; the longer cue interval controls at 310–311; the lower-contrast judgment control at 311–312. Orientation accuracy improved within the dynamic range at both tested spatial frequencies in the low-contrast experiment. In the high-contrast experiment it improved at 4 cycles per degree, while performance at 2 cycles per degree was already at ceiling.

The main text therefore distinguishes evidence of improved discrimination from the appearance judgments used to infer changed apparent contrast. Patrick Cavanagh, Sheng He, and James Intriligator, "Attentional Resolution: The Grain and Locus of Visual Awareness," in *Neuronal Basis and Psychological Aspects of Consciousness*, ed. C. Taddei-Ferretti and C. Musio (Singapore: World Scientific, 1999), 41–52, esp. 41–43, argues that attention has a coarser spatial grain than early vision and thereby limits which registered detail reaches visual awareness. The interpretation remains contested: an apparent-contrast result can reflect represented contrast, representational determinacy, or post-perceptual decision effects. The intentionalist cannot infer the content from the reported feel alone; the empirical work constrains which independently specified representational variable changed. This intervention changes attention and present processing; it does not manipulate content-fixing history while preserving current structure. Its results therefore do not discriminate historically wide character from structural representationalism.

**12.** Frank Jackson, "Epiphenomenal Qualia," *Philosophical Quarterly* 32 (1982): 127–136, esp. 130, and "What Mary Didn't Know," *Journal of Philosophy* 83 (1986): 291–295, esp. 291–94. Jackson's case has generated an enormous secondary literature; the canonical collection is Peter Ludlow, Yujin Nagasawa, and Daniel Stoljar, eds., *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument* (Cambridge, MA: MIT Press, 2004).

**13.** The reply pressed here — that Mary acquires a new representational route to a property she already knew, underwriting a new perceptual demonstrative thought ("*this* is what red looks like") rather than any new fact — is developed in full, and set against its neighboring rivals (including the ability hypothesis), in Chapter 8. For a related anti-dualist route that grants Mary new subjective facts while denying that all facts must be physical, see Tim Crane, "Subjective Facts," in *Real Metaphysics: Essays in Honour of D. H. Mellor*, ed. Hallvard Lillehammer and Gonzalo Rodríguez-Pereyra (London: Routledge, 2003), 68–83, esp. secs. 2 and 6.

**14.** Jackson's later reassessment appears in "Mind and Illusion," in *Minds and Persons*, ed. Anthony O'Hear (Cambridge: Cambridge University Press, 2003), 251–272, esp. 251–52; Chapter 8 walks that change of mind as the chapter's human spine.

**14a.** For an influential formulation, see Sydney Shoemaker, "The Inverted Spectrum," *Journal of Philosophy* 79, no. 7 (1982): 357–381. The representationalist reply through asymmetries in color space is developed in Michael Tye, *Ten Problems of Consciousness* (Cambridge, MA: MIT Press, 1995), ch. 7. The geometry weakens the inference from a describable inversion to a possible undetectable difference; it does not prove inversion impossible. Block, "Mental Paint," explicitly considers possible creatures with the asymmetries removed. Human color geometry cannot exclude their inversion by itself. Appendix B.5 develops the relational-structure reply and its limits.

**15.** The case originates in Ned Block, "Inverted Earth," *Philosophical Perspectives* 4 (1990), esp. 62–64, and is restated and pressed in Block, "Mental Paint" (cited in n. 5 above). Block intends it as the scenario externalist representationalism cannot accommodate: by holding the subject's phenomenal character fixed (the inverting lenses) while the wide content of his color states inverts (the move to a complementary-reflectance environment, followed by re-calibration), the

case pulls phenomenal character and representational content apart — two states identical in felt character, opposite in what they represent. It is the diachronic, cross-world strengthening of the static inverted-spectrum case of Section 6.4.2 (developed in Appendix B.5): where the inverted spectrum can be met by the integrated relational structure of color experience, Inverted Earth leaves that structure untouched and dissociates content from character at the level of world-involving wide content, which is the level the identity claim occupies. The construction separates the theories' verdicts, not independently measured phenomenal characters. Matching proximal stimulation does not by itself establish phenomenal constancy; semantic re-anchoring does not by itself establish phenomenal change. The book adopts the latter verdict through its wide identity, with the lower evidential status specified in Sections 6.4.3 and 6.5.3.

**16.** John R. Searle, *Seeing Things as They Are: A Theory of Perception* (Oxford: Oxford University Press, 2015), esp. chs. 1 and 3, develops the diagnosis at book length; see also Searle, "The Philosophy of Perception and the Bad Argument," in *Wirklichkeit oder Konstruktion? Sprachtheoretische und interdisziplinäre Aspekte einer brisanten Alternative*, ed. Ekkehard Felder and Andreas Gardt (Berlin: De Gruyter, 2018), 66–76. The Bad Argument is not merely the inference from sometimes noticing how things seem to always noticing only appearances. It combines the possibility of hallucination with a common analysis of veridical and hallucinatory experience, then mistakes the content common to the cases for a common object of awareness. Searle's own alternative gives perceptual content causally self-reflexive, *de re* conditions of satisfaction: the experience presents its object directly by requiring that very object to cause that very experience. The book borrows the diagnosis and the compatibility of intentionalism with direct realism, not this machinery. Its abstract content concerns property types; an appropriate current causal relation within a causal-historical organization fixes the token object, a debt Chapter 15 pays. Tim Crane and Craig French classify intentionalism as a direct-realist theory on this causal understanding: direct perceptual representation plus the appropriate non-deviant causal relation to the represented object constitutes seeing, while the content common to perception and hallucination supplies no common object of awareness. See Crane and French, "The Problem of Perception," *Stanford Encyclopedia of Philosophy* (Fall 2021 edition), secs. 3.3.3–3.3.6.

**17.** The terminological tangle is documented in James Genone, "Recent Work on Naïve Realism," *American Philosophical Quarterly* 53, no. 1 (2016): 1–25, esp. sec. 2, which separates direct realism (the denial that we perceive the world by perceiving something else first), representationalism (perceptual experience has representational content), and naïve realism (perceived objects partly constitute experience) as three distinct theses too often run together under one label. Direct realism is the genus; representationalism and naïve realism are rival species of it, agreeing that perception is direct and disagreeing about what makes it so. Charles Travis's prior challenge appears in "The Silence of the Senses," *Mind* 113, no. 449 (2004): 57–94, esp. 57–60 and 73–78: perceptual experience, as such, need not select one truth-evaluable face value rather than countless compatible descriptions. For John Campbell's argument that object-involving acquaintance grounds demonstrative reference, and for its relation to disjunctivism, see Matthew Soteriou, "The Disjunctive Theory of Perception," *Stanford Encyclopedia of Philosophy*, substantive revision July 28, 2026, secs. 3.5–3.6, which reconstructs the argument and surveys Campbell's primary treatments. The main text answers both at the hinge: content must receive independent accuracy

conditions and token reference rather than inherit either from the phenomenal report or from an already mastered demonstrative.

**18.** Bill Brewer, "Perception and Content," *European Journal of Philosophy* 14, no. 2 (2006): 165–181, esp. 165–170, rejects the content view he had previously defended because its truth-evaluable generality and possibility of falsity fail, in his judgment, to capture perception as direct conscious access to the mind-independent objects themselves. He develops the positive Object View in *Perception and Its Objects* (Oxford: Oxford University Press, 2011). The related constitutive reading is developed in M. G. F. Martin, "The Transparency of Experience," *Mind & Language* 17 (2002): 376–425, https://doi.org/10.1111/1468-0017.00205. The disjunctivist defense and the indiscriminability-based treatment of hallucination appear in Martin, "The Limits of Self-Awareness," *Philosophical Studies* 120 (2004): 37–89, esp. 38–41 and 67–72, and "On Being Alienated," in *Perceptual Experience*, ed. Tamar Szabó Gendler and John Hawthorne (Oxford: Oxford University Press, 2006), 354–410. The pressure point is hallucination: because the perceptual relation is genuine, both relata must exist, so a veridical perception and a matching hallucination cannot share their fundamental nature. Martin accepts that the cases may be introspectively indiscriminable; he denies that this epistemic match establishes positive phenomenal sameness. The representationalist must therefore defend shared content abductively, through its explanatory work, rather than treat it as common ground. For the distinctions among fundamental-kind, content, phenomenal, and epistemological disjunctivisms, and for relational views that allow more commonality than Martin does, see Soteriou, "The Disjunctive Theory of Perception," secs. 2–5.

**18a.** Hilary Putnam, "The Meaning of 'Meaning,'" in *Language, Mind, and Knowledge*, ed. Keith Gunderson, *Minnesota Studies in the Philosophy of Science* 7 (Minneapolis: University of Minnesota Press, 1975), 131–193, at 141, allows that Oscar and his counterpart "were exact duplicates in appearance, feelings, thoughts, interior monologue, etc." while the extension of "water" differs; the argument concerns extension and "meaning" in the intuitive sense, and it makes no claim about phenomenal character. The two-layer reading of perceptual content adopted in the text follows the structure of Michael Tye, *Consciousness Revisited: Materialism without Phenomenal Concepts* (Cambridge, MA: MIT Press, 2009). Tye there embraces the Singular (When Filled) Thesis, on which veridical experience has a singular content into which the seen object enters and hallucination a gappy content of the same type (ch. 4, sec. 4.7, 94). He locates the commonality in the "non-object involving" properties that enter both contents (ch. 5, 115–116), and he concludes that the phenomenal character of a perceptual experience "consists in, and is no more than, the complex of qualities the experience represents" (ch. 5, 119). The book takes over that division of labor. It does not take over Tye's further treatment of phenomenal character as a complex of external qualities, and it keeps its own identity claim at the level of the presented properties under a perceptual mode (Section 6.5.3). Two costs follow. First, cognitive-phenomenology and rich-content views, on which perception can present kind properties such as *pine tree* or *water*, must be answered by showing that the relevant expertise contrasts change the presented low-level properties or their attentional organization; Susanna Siegel, *The Contents of Visual Experience* (New York: Oxford University Press, 2010), ch. 4, "Kinds," 99–116, presses the strongest case. Second, object-involving character, as relationalists hold (n. 18), would make the twins differ phenomenally, which Putnam's stipulation does not decide. Because the property layer's content is itself fixed by world-

involving history and use, the view remains externalist about character: Inverted Earth (n. 15) and Swampman (Chapter 15) are the cases that separate it from narrow accounts. Twin Earth is not. The proposed boundary distinguishes nonconceptual property attribution from singular object reference and conceptual classification by their explanatory roles. Independent discrimination, constancy, learning, and correction constrain the first; mere contact with a substance does not establish representation of its chemical kind. This is a constraint on content attribution, not a proof that historical property reference constitutes character. A structural account can accept these distinctions while assigning phenomenal work to the present profile alone. Inverted Earth and Swampman discriminate the theories' commitments; neither provides independent evidence selecting the wide verdict. The historical extension accordingly carries less achieved warrant than the shared intentionalist explanation of ordinary perceptual contrasts. Tye, *Consciousness, Color, and Content*, sec. 6.4, 139–140, explicitly distinguishes a perceptible property from an imperceptible correlate and proposes asymmetric counterfactual dependence to exclude the latter from sensory content. That proposal supports the need for a principled content boundary; it does not establish the book's historical account of that boundary or its phenomenal consequences. Byrne, "Intentionalism Defended," *Philosophical Review* 110 (2001): 199–240, sec. 2, formulates a determination thesis while separating it from reductive and psychosemantic commitments. Byrne and Tye, "Qualia ain't in the head," *Noûs* 40 (2006): 241–255, secs. 1.1.1 and 2, distinguish neural dependence from internal sufficiency and externalist intentionalism from a completed reductive content theory. These distinctions constrain the present comparison; none independently selects its historical extension.

**18b.** For reflectance realism, see Alex Byrne and David R. Hilbert, "Colors and Reflectances," in *Readings on Color, Volume 1: The Philosophy of Color*, ed. Byrne and Hilbert (Cambridge, MA: MIT Press, 1997), 263–288, and "Color Realism and Color Science," *Behavioral and Brain Sciences* 26 (2003): 3–21. C. L. Hardin, *Color for Philosophers: Unweaving the Rainbow* (Indianapolis: Hackett, 1988), develops the error-theoretic and observer-variation challenge; John Campbell, "A Simple View of Colour," in *Reality, Representation, and Projection*, ed. John Haldane and Crispin Wright (New York: Oxford University Press, 1993), 257–268, defends a primitivist alternative.

**19.** The debt Section 6.5 marks is the burden of *aboutness*: the directness defended there runs entirely through the experience's content being satisfied only by its object, so the whole apparatus leans on the further claim that this aboutness is a natural fact rather than a spook. Strong representationalism cannot help itself to directness on credit. The contrast is between two ways of treating intentionality. Searle holds that intentionality is intrinsic to certain biological systems and not further reducible: "that property of many mental states and events by which they are directed at or about or of objects and states of affairs in the world" (John R. Searle, *Intentionality: An Essay in the Philosophy of Mind* [Cambridge: Cambridge University Press, 1983], 1), a property he takes to require the right causal powers, those of brains or of anything, organic or not, that duplicated them, but never the running of a formal program alone. Whether other hardware could have those powers he leaves, in the *Minds, Brains, and Programs* exchange, an empirical question. He also takes intentionality to resist explanation in more basic terms and offers no reductive account of what perceptual aboutness consists in. This book takes the opposing, naturalizing line: aboutness is a physical relation cashable through selection, learning, tracking, producer-consumer

organization, and world-involving history and use. The commitment is introduced in Chapter 1 (intentionality as world-involving, content fixed by a system's history of engagement rather than by intrinsic inner stuff) and the case for it is made in Chapter 15, which develops the tracking-and-function tradition of Fred Dretske, *Knowledge and the Flow of Information* (Cambridge, MA: MIT Press, 1981), esp. ch. 7, and *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995), esp. ch. 1, and Ruth Garrett Millikan, *Language, Thought, and Other Biological Categories* (Cambridge, MA: MIT Press, 1984), esp. chs. 2 and 6. The two strands differ in emphasis — Dretske begins from an information-theoretic notion of indication and recruits teleology to fix the content where mere correlation underdetermines it, while Millikan begins from etiological biological function — but both cash representation in terms of what a state has the function of indicating, and Chapter 15 draws on both, with the frog's fly-detector and the misrepresentation it makes possible as the test case. The point is dialectical, not doctrinal: Searle's cure for the Bad Argument can be adopted without adopting his primitivism about intentionality, and indeed must be, since a directness that bottoms out in unexplained aboutness has relocated the mystery rather than removed it — which is also where this book's intentionalism parts from Searle's, even as it borrows his cure. Chapter 15 also meets the standard objection to any history-based account of content — Davidson's Swampman (Donald Davidson, "Knowing One's Own Mind," *Proceedings and Addresses of the American Philosophical Association* 60 [1987]: 441–458, esp. 443–44), the molecule-for-molecule duplicate with no past — by accepting absence rather than disguising it as indeterminacy: at the first instant the duplicate lacks any content-relevant history and therefore lacks content and, on this chapter's identity claim, phenomenal character; consequential engagement can begin the learning history from which both later arise. Chapter 15 develops the general naturalizing model and its worked applications. Evaluative pain content, mood's diffuse accuracy conditions, and the warrant for experience-directed relief remain substantive limits on its completeness; locating the debt does not discharge those applications.

**19a.** Dennett develops the Multiple Drafts model against the Cartesian Theater in *Consciousness Explained* (Boston: Little, Brown, 1991), chs. 5–6. The chapter takes the distributed architecture as a positive resource. Dennett's qualia eliminativism denies additional intrinsic qualitative properties, not the reality of pain or of content-bearing experience; see ch. 12 and the Appendix. The house identity claim remains a distinct positive commitment rather than following from that shared denial.

**20.** Michael Tye, *Ten Problems of Consciousness* (Cambridge, MA: MIT Press, 1995), ch. 5, sec. 5.2 (quotation at 162), defines poised content as content "poised for use in the formation of beliefs and/or desires." The book uses PANIC as a historical framework, not as Tye's unrevised current position: Tye later calls PANIC his "old view (1995)" and rejects its equivalence to his newer externalist account in "Yes, Phenomenal Character Really Is Out There in the World," *Philosophy and Phenomenological Research* 91, no. 2 (2015): 483–488, secs. 11–17, https://doi.org/10.1111/phpr.12167. Daniel C. Dennett's Multiple Drafts model rejects a finishing place where content becomes conscious and instead tracks distributed processes whose effects on memory, report, reasoning, and action give some contents greater "fame"; *Consciousness Explained* (Boston: Little, Brown, 1991), chs. 5–6. The present synthesis treats that architecture as evidence for a distributed access-and-control profile while rejecting the inference that phenomenal character should be eliminated.

The four PANIC conditions distinguish phenomenally conscious states from unconscious representations, conceptual attitudes, and mere sensory registrations, but Poise means availability to a cognitive consumer, not inevitable report or judgment. Ned Block, "Consciousness, Accessibility, and the Mesh between Psychology and Neuroscience," *Behavioral and Brain Sciences* 30 (2007): 481–499, secs. 3–4 and 9, distinguishes actual report, cognitive accessibility, and the constitutive role of access. His discussion of Sperling's partial-report results allows a cued row to be reported while arguing that phenomenology exceeds cognitive accessibility. The book's selective-availability interpretation explains why limited actual report alone does not decide the issue; it does not refute Block's stronger constitutive challenge. The Nonconceptual condition invites a Sellarsian objection worth heading off. Tye defines "nonconceptual" by concept possession: the general features entering the content "need not be ones for which their subjects possess matching concepts" (*Ten Problems*, sec. 5.2); fine-grained color discrimination is his motivating example, not the definition. The condition does not posit a brute sensory item licensing belief from outside the space of reasons; compare Wilfrid Sellars, "Empiricism and the Philosophy of Mind" (1956). Fine-grained content can stand available for conceptual uptake without already being conceptual or automatically consumed. See Michael Tye, "Nonconceptual Content, Richness, and Fineness of Grain," in *Perceptual Experience*, ed. Tamar Szabó Gendler and John Hawthorne (Oxford: Oxford University Press, 2006), and contrast John McDowell, *Mind and World* (Cambridge, MA: Harvard University Press, 1994).

**21c.** The table includes only identity assignments. Substance dualism and property dualism refuse the identification; Chapter 9 meets that refusal. Illusionism denies the phenomenal property rather than identifying it (Keith Frankish, "Illusionism as a Theory of Consciousness," *Journal of Consciousness Studies* 23, nos. 11–12 [2016]). Behaviorism offers a disposition; Chapter 20 treats it as the analysis functionalism replaced. U. T. Place's 1956 paper, cited in n. 21a, frames the hypothesis as the is of composition, on the model of a cloud and its droplets, not as a definitional identity. J. J. C. Smart's 1959 paper, cited there with it, is the strict type identity. Donald Davidson, "Mental Events" (1970), in *Experience and Theory*, ed. Lawrence Foster and J. W. Swanson, 79–101, reprinted in *Essays on Actions and Events*, holds that mental events are identical with physical events, taking events as unrepeatable dated individuals, while denying strict psychophysical laws. Chapter 20, notes 4 and 5, separates Hilary Putnam's identification of the state with the machine-table role from David Lewis's identification of the state with the occupant. Chapter 10 distinguishes the identity reading of Russellian monism, on which the categorical bases of physical roles are phenomenal, from John Heil's powerful-quality identity, one property under a qualitative description and a causal one, and from constitutive panqualityism, which constitutes a whole rather than identifying the feel with its basis. Galen Strawson's real physicalism reaches that Russellian drop by another road, the intrinsic nature of matter: experience is wholly physical, and wholly non-experiential foundations could not yield it.

**21d.** The higher-order theory met in n. 7b is a theory of state consciousness, not an identification of phenomenal character with the higher-order thought. Integrated information theory: Giulio Tononi and Christof Koch, "Consciousness: Here, There and Everywhere?," *Philosophical Transactions of the Royal Society B* 370 (2015): 20140167; Giulio Tononi, Melanie Boly, Marcello Massimini, and Christof Koch, "Integrated Information Theory: From Consciousness to Its Phys-

ical Substrate," *Nature Reviews Neuroscience* 17 (2016): 450–461. Alva Noë, *Action in Perception* (Cambridge, MA: MIT Press, 2004), has sensorimotor skill constitute the experience and sensorimotor knowledge give it content. Place's composition, n. 21c, identifies the sensation with the brain process. Noë's constitution leaves the skill and the experience as two relata unless a further identity is stated. The sorter is the number of relata, not the verb. Phenomenal intentionality, on which phenomenal character is the source of intentional content: Angela Mendelovici, *The Phenomenal Basis of Intentionality* (New York: Oxford University Press, 2018), esp. the chapter "The Phenomenal Intentionality Theory" (DOI 10.1093/oso/9780190863807.003.0005).

**21.** Michael Tye gives an explicit identity formulation: "phenomenal character is one and the same as representational content that meets certain further conditions" (Michael Tye, *Consciousness, Color, and Content* [Cambridge, MA: MIT Press, 2000], ch. 3). This establishes the form of the claim, not agreement about its precise relatum. The book's content-under-mode profile belongs with impure rather than pure intentionalism: mode can contribute without being another property represented as belonging to the object. Compare Crane's distinction between intentional content and mode, cited in n. 10; the book does not thereby adopt his whole account. Nor does it identify its profile with Tye's later complex of external qualities (n. 18a). "Consists in" names numerical identity in the main text, not the weaker constitution relation exemplified by Lynne Rudder Baker, *Persons and Bodies: A Constitution View* (Cambridge: Cambridge University Press, 2000). Naming the distinction does not refute the modal rival. Chapters 9 and 11 test whether conceivability and explanatory difference supply a genuine separation between character and the complete, independently specified intentional profile.

**21a.** For theoretical identification, U. T. Place, "Is Consciousness a Brain Process?," *British Journal of Psychology* 47 (1956): 44–50, and J. J. C. Smart, "Sensations and Brain Processes," *Philosophical Review* 68 (1959): 141–156. The water/$H_2O$ and heat/molecular-motion analogies serve here as heuristics for the logical form of an *a posteriori* identity — conceptual difference without ontological difference — not as historical demonstrations that explanatory convergence entails identity. The adoption in the text is abductive and defeasible, per the conditions it posts.

**21b.** Constitution's paradigm cases carry independently motivated differences between the relata — persistence conditions, modal profile — that support both the nonidentity and the direction of dependence. Lynne Rudder Baker, *Persons and Bodies: A Constitution View* (Cambridge: Cambridge University Press, 2000), builds the view on differences of exactly this kind; Tye deploys the statue and the clay the same way in *Consciousness and Persons: Unity and Identity* (Cambridge, MA: MIT Press, 2003). The rival's most careful contemporary form replaces constitution with grounding, an asymmetric and hyperintensional dependence built to survive agreement in causal and modal profile: see Kit Fine, "Guide to Ground," in *Metaphysical Grounding: Understanding the Structure of Reality*, ed. Fabrice Correia and Benjamin Schnieder (Cambridge: Cambridge University Press, 2012), 37–80. The text claims no refutation of grounding-theoretic nonidentity; it observes that no independent motivation for the asymmetry has been supplied in this case, and conditions the identity claim's adoption on that absence. Kim, *Physicalism, or Something Near Enough*, ch. 5, 121–148, argues that identities do not explain correlations (131ff.); the view that they make the

demand for an explanation ‘wrongheaded’ is Block and Stalnaker’s, after Smart and Feigl, which Kim grants only conditionally (138–140).

**22.** David Chalmers, “Facing Up to the Problem of Consciousness,” *Journal of Consciousness Studies* 2 (1995): 200–219, esp. 200–03, and *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press, 1996). The identity claim treats the explanatory demand as epistemic rather than as evidence for an ontological remainder: if a conscious state’s phenomenal character and its complete world-presenting profile are identical, no causal bridge connects two properties in nature, though the conceptual question of how the identity can be understood remains. Chapter 11 engages that distinction directly.

## Chapter 7: The Hard Cases

> *"Pains are experiences that represent tissue damage in bodily parts."*
>
> — Michael Tye, "The Experience of Emotion: An Intentionalist Theory"

---

CHAPTER OVERVIEW

A toothache at three in the morning and a gray mood over the day seem to point nowhere. I take them as the strongest tests of the claim that felt character consists in what a state presents and how it presents it, since explaining where pain points still leaves the question of why it hurts. Phantom pain gives the bodily-address account a case to explain, while pain asymbolia reveals a split between location and distress without deciding what that split means. I favor a view on which moods present the body and the day under a felt cast, but a feeling that causes those changes remains a serious rival. These cases keep the identity claim in play; they do not extend the book's achieved explanation from perception to pain's badness or mood. The universal thesis remains my view, but its support is uneven: the affective account still has to earn the content it invokes.

### 7.1 The Intentionality Objection

The identity claim hands its opponents a clean target. Produce a state that unmistakably feels like something and represents nothing, and the claim collapses. Unlike the elaborate thought experiments to come, this objection needs no philosophical machinery; it asks only that you stub your toe.

> *Look,* it says, *the representationalist has told a lovely story about seeing tomatoes. Vision is about the world; fine. But not all experience is vision, and the cases you have quietly set to one side are the ones that matter. Take a toothache at three in the morning. What does it represent? Nothing. It does not present a tomato or a tree or any worldly object; it simply hurts, sitting there in the jaw as a raw felt quality, a throb that is about as much "about" something as a cramp is an argument. Take a mood — the flat gray dread that is on you the moment you wake, before the coffee, before the news, attached to nothing, pointed at nothing, and yet unmistakably felt. These states have phenomenal character in abundance. They have representational content nowhere. They are the felt-quality-without-content the identity claim says cannot exist. The claim is not refined by these cases. It is refuted by them.*

The objection identifies the right test. A toothache need not present an external object as vision does to represent something: it may present *the body*. But that reply only begins the argument.

Pain's location, its hurting, and a mood's diffuse hold on the day each need an account; giving one of them content cannot settle the others.

### 7.2 The Inner-Sensation Picture, and the Phantom Limb

Start with the account of pain so obvious it scarcely feels like a theory. On this picture, a pain is a private hurt—a felt thing, a quale—residing wherever the mind resides, which Cartesian inheritance places somewhere behind the eyes. The toe merely seems to hurt, a geographical illusion staged by the nervous system. Signals travel upward; the hurting happens in the inner theater; the mind projects it back down as a useful fiction.

Consider phantom-limb pain. A man wakes in a hospital bed three weeks after the accident, and his right foot hurts — not vaguely: it burns, it throbs, the toes cramp as though someone were bending them back. The foot was amputated below the knee. He knows this. He can see the bandaged stump where the leg now ends. And the toes he no longer has go on aching, vividly, as if a hot wire were laid against them. The neuroscientist V. S. Ramachandran documents them by the ward-full; he has spent a career turning these uncanny cases into windows onto how the brain builds the body, and once relieved a phantom's clenched-fist cramp with nothing but a mirror and a cardboard box. They are not exotic. Most amputees report phantom sensations, and a large fraction of those are painful.[1]

The inner-sensation picture cannot place phantom pain in missing toes, so it puts the pain in the brain and treats its felt location as projection.

If the phantom's pain is really in his head and only seems to be in his toes, then the same must hold for *every* pain, since the phantom case differs from the ordinary case only in the absence of the limb, not in the character of the experience. So when a child snatches her hand back from the stove and wails that her *finger* hurts, the strict inner-sensation theory must treat the hurt as intracranial and the finger as its merely projected address. A sophisticated defender can accept that result. She can say that ordinary pain talk reports the projection rather than making a naive anatomical claim. Fair enough. Both accounts need the brain's machinery for locating bodily trouble; they differ over what that machinery delivers. On the projection account, it delivers two things: a private hurt, and a location assigned to it afterward. On the representational account, location belongs to what the state presents from the start. The issue turns on economy and unification, not on convicting the child of bad grammar.

Chapter 5′s vehicle/object distinction removes one reason for relocating the hurt: the neural process need not occupy the place the experience presents as hurting. That answers the location puzzle. It does not yet answer a theorist who grants the bodily representation and holds that its painfulness consists in a further intrinsic quality. She need posit no inner spectator. Her objection waits in Section 7.4.

### 7.3 Pain as the Body's Alarm

The alternative explains ordinary and phantom pain through one bodily-address relation.[2]

A pain experience *represents a condition of the body*. Provisionally, its content presents an actual or impending threat to bodily integrity at a represented location as bad for you, under a bodily-

affective mode: *something threatens this part of you, and it matters.* The "here" belongs to the content. To have a pain in the toe is to undergo an experience that presents a threat at the toe. A headache or toothache can instead present a condition within the head; nothing in this account requires an object outside the skull. The distinction concerns the neural vehicle and the represented bodily condition, wherever that condition lies. Chapter 15 must earn the naturalistic details of this content rather than letting this chapter adjust them case by case.

Think of pain as a notice posted at a bodily site: *threat here, attend now.* Like any notice, it can be posted accurately or in error.

In the ordinary case, you stub your toe and the pain accurately presents a threat there. That is why we say, without strain or error, that the toe hurts.

In the phantom case, the same kind of experience presents a threat at absent toes and so misrepresents. The phantom and the stubbed toe share a bodily address; they differ in accuracy, like the bent stick and the straight one. There is no inner item to relocate — only a representation that sometimes gets the body wrong. In pain, "outward" reaches only as far as the body.[3]

The alarm, in short, went on sounding in a wing of the building that no longer stood. The wiring did not know.

Understanding why it persists does not make living with it easier. The mistake concerns the represented foot, not the reality of the man's suffering.

A tickle near the shoulder presents an arrangement: direction, distance, and boundary, under a bodily mode. It also concerns a region, the place that would make the presentation accurate. Move the finger and the arrangement changes. The felt relations change, and the processing that carries them changes with them. That is why the character changes. It does not show that character changes when only the region does.

The phantom makes the split plain. The limb is gone. The arrangement remains. You feel the body as reaching farther than it does, not a hurt hovering off the stump. History can explain why that missing limb is still the target. It does not supply the distance.

A structural rival, Chapter 6′s alternative on which character consists in an experience's present structure while history fixes only what it refers to, asked whether history supplies the apparent geometry, or only an address for a geometry already apparent. Here the arrangement belongs with what the experience presents now. The address belongs with what would make that presentation right or wrong. Bodily location does not prove Chapter 6′s wide identity, on which history, by fixing which properties an experience presents, helps make the feel itself. On both views, the particular part referred to settles accuracy, not feel. It shows where the two jobs come apart.

### 7.4 But Pain *Hurts*

The objection now reaches the genuinely hard point.

> *Fine,* the objection says. *Grant that pain has a locative, bodily content — that it represents the toe, the tooth, the back. You have explained the*

> *part of pain that points somewhere. You have not touched the part that matters. Because pain does not merely* locate*; it* hurts*. There is something awful about it, a felt badness, and surely* that* — the sheer hurting — is not a representation of anything. The hurting is the inner quale the representational account keeps promising to reach and never reaches. You have re-described the address on the envelope and said nothing about why the letter is a scream.*

Repeating that pain represents the body will not answer this objection. Consider one aching ankle. The ache locates the disturbance *there*, presents it as *bad for you*, and presses you to *keep weight off it*. Those contributions can differ: you might recognize the location yet feel little distress, or feel the distress while deciding that walking to safety matters more. Locating, evaluating, and directing action do different work. An evaluation can misrepresent; a command can go unfulfilled even when acting on it would help. The account identifies hurting with the bodily-affective presentation of badness, not with obeying the instruction or coolly judging that the ankle needs care.

Two terms must stay apart. *Badness* belongs to what the ache presents: the ankle's condition as bad for you, which the ache can get right or wrong. The *bodily-affective mode* names how the ache presents it: felt from within, and pressing. Chapter 6 counts both in the profile that phenomenal character consists in. Neither is mental paint. Philosophers call this an *impure* intentionalism, because the manner of presenting counts along with what is presented. A cool judgment that the ankle is in bad shape can present the same badness without that mode, and it does not hurt. On this reading the phantom misrepresents twice: no foot is threatened, and no foot is in a condition bad for him. His suffering stays real, because hurting consists in the presentation, not in its accuracy.

David Bain develops the *evaluativist* account: pain presents a bodily disturbance as bad for the subject, and its unpleasantness consists in that evaluative content.[4] Colin Klein's *imperativist* account instead treats pain as an order to protect the body. It captures the ankle's demand without making that demand a statement. The view defended here keeps evaluative, locative content because it needs an account of how pain gets the body right or wrong. Calling the presentation evaluative identifies the proposed answer; it does not prove that nothing remains beyond it.

Hilla Jacobson asked why we so often try to silence the messenger, and Matthew Kinakin presses the ordinary painkiller harder. If pain presents the bodily condition as bad, why aim the drug at the experience when the condition remains? And what warrants the judgment *I cannot stand how this feels*? Part of the answer concerns gain control: an alarm can outlive its usefulness, consuming attention, sleep, and coordinated action without improving protection. Brian Cutter and Michael Tye distinguish the reason to treat a represented threat from the reason to want the signaling state gone. The messenger belongs to the control system too.[5]

The painkiller presses further than wasted attention or lost sleep. Imagine a brief ache while you sit safely, with no task to finish and no further harm at stake. You still have reason to want the hurting to stop. A reason to protect the ankle and a reason to end this experience therefore need separate accounts. Saying that undergoing the state gives access to its badness names the access

we need to explain. It does not show how a profile fixed without appeal to the hurt supplies that warrant. Kinakin's challenge remains unanswered here. That leaves a gap in this account, not a reason to doubt the suffering or the reason for relief.

Why should even evaluative content *hurt*? A dashboard can flag damage and stay as cool as its glass. Chapter 6′s proposal concerns a bodily-affective presentation in a conscious subject, not a warning symbol. But calling a mode *affective* cannot supply the missing account of badness. The book's better-supported cases concern how things look: blur, attention, and the properties a visual field presents. They support a general identity thesis without completing its application to pain. Section 10.6 asks whether a further phenomenal ingredient earns its keep. Failure to earn that ingredient would still not turn the dashboard into an explanation of hurting.

In **pain asymbolia**, usually associated with damage involving the insular cortex, patients respond to noxious stimulation with strikingly little distress or avoidance.[6] Reports in the classic literature describe preserved discrimination and location alongside flattened affective and motivational response. Whether those subjects still feel *pain* without painfulness, rather than registering noxious stimulation without pain, has become a live dispute. That taxonomic question matters. The chapter cannot settle it by borrowing the word *pain* from the syndrome's name.

Under the standard interpretation, bodily location and affect come apart. That fits the distinction the ankle case introduced, with sensory discrimination and felt badness available to different forms of disruption.

The quale theorist therefore has a real reply. She can say that asymbolia removes the patient's aversive reaction while leaving sensory content and intrinsic sensory character intact; Grahek's phrase *reactive dissociation* leaves room for that reading. The representationalist answers that the missing dimension behaves like an evaluation, altering attention, significance, and action together. Yet a theory of content plus reaction can explain that pattern too. Asymbolia identifies dimensions an adequate theory must distinguish. The preference for identifying affect with evaluative presentation rests on the wider case in Chapter 6, not on a clinical result that already chose our metaphysics.

The toothache has a bodily address the account can explain. Its badness, and our grounds for wanting that very experience to end, remain harder. I defend the same identity for pain as for vision, but locating the tooth has not yet explained the hurt.

### 7.5 The Intransitive Objection

Murat Aydede presses the pain objection as a fellow naturalist demanding that perceptual and strong representational theories explain pain's peculiar phenomenology rather than merely rename it.[7] The surface grammar makes the trouble vivid. When I say *I see a tree*, the verb takes an object, something my seeing concerns and can get right or wrong. But *I have a pain in my foot* reads less like *I see a tree* than like *I have a bruise on my foot*. The pain seems merely located, sitting at a place, not directed at an object the way perception reaches the world. Aydede's stronger point survives any quick grammatical reply: the representationalist must supply independent grounds for accuracy conditions instead of declaring pain perceptual because phantom cases make that description convenient.

The reply grants the grammar and makes a comparative case. The locative "in" of *a pain in my foot* behaves unlike the bare location of *a bruise on my foot*. Pain directs protection toward a represented region; ordinary, referred, and phantom pain can conflict with touch, vision, and the body's actual condition; and one bodily-address relation explains both the matches and the mismatches. The sensation theorist can call location a projected feature, but then projection must reproduce the same pattern while leaving the private hurt itself without an address. She can instead call location an intrinsic feature of the sensation. Then she must say why that feature counts as wrong in referred pain, felt in the arm while the trouble lies in the heart, and why sight and touch can correct the urge to nurse the arm. A feature that can be wrong about the body already does an address's work. The representational account builds the address into the state from the start.

The phantom-bruise contrast makes that comparison memorable without proving it by itself. A bruise merely occupies its place. It cannot disagree with an empty foot. A pain can present the absent foot as threatened even while vision and belief insist that no foot remains. You cannot have a phantom bruise. You can have a phantom pain. That difference gives us an abductive clue: pain behaves like a state answerable to the body it presents. Chapter 15′s proper-function account must supply the final warrant for calling that bodily address accurate or inaccurate. For now, the representational reading unifies more of the pattern than projection does, with fewer posits: no hurt that first exists without a place and then receives one. Against intrinsic location, it rests on the reply just given.

Pain's bodily address gives the contentless reading something to explain. That leaves open whether hurting itself consists in evaluative content under a bodily-affective mode. Asymbolia helps separate the questions; it does not decide between the answers.

### 7.6 Moods, the Last Refuge

Pain at least happens *somewhere*: the tooth, the toe, the back. Moods need not offer even that foothold. You wake in a bad mood. Nothing has happened. The coffee has not burned, the news has not arrived, the cat has not yet knocked anything off the dresser. The day sits in front of you, blank, and already it feels wrong. Pressed to say what the mood is *about*, you would shrug; pressed harder you might gesture at the weather, or last night's dream, or the slow gray weight of February, and none of it would quite fit. The mood arrived uninvited and seems to point at nothing whatever. Here, surely, is phenomenal character with no object at all.[8]

The temptation rests partly on grammar. We expect a representational state to take a bounded object: *I see the lemon. I fear the dog. I remember the kitchen.* Free-floating dread seems to lack the required noun.

The missing noun leaves several possibilities. Three intentionalist accounts give the mood content concerning:

1. a global condition of the body;
2. the surrounding field taken as a whole; or
3. an affective property represented without any bearer at all.

Watch what a mood actually does. Depression does not show up as one more item in consciousness, sitting in a corner next to the lemon and the kitchen. It shows up as the way *everything else* shows up. The coffee tastes faintly of nothing; the conversations feel thin and far away; the future tilts subtly downward; the kitchen you have stood in ten thousand times looks, this morning, like a place to be gotten through. Nothing in particular looks bad. The whole field has acquired a negative cast, a kind of weather. Matthew Ratcliffe describes such states as bodily background orientations that alter salience, possibility, familiarity, and the shape of action itself.[9] He does not thereby join the representationalist camp. He supplies the phenomenological bridge: mood can reach beyond a discrete object without collapsing into an inner patch of weather.

The bodily reading starts from anxiety's tight chest, depression's weight, elation's lift, and what trained attention often finds when it turns toward a mood. That account explains the mood's somatic grip and asks no mysterious object to carry it. But body alone can underdescribe the way a mood transforms the future, other people, and the space of possible action. The field reading catches that reach. It treats the mood as presenting the world or situation under a diffuse affective mode, a wash of threat or warmth or wrongness over the scene. Its cost lies in the phrase *the world as a whole*: unless Chapter 15 can supply determinate accuracy conditions, "everything" threatens to become a blank check rather than content.

Angela Mendelovici offers the third account. Some moods, she argues, present affective properties *unbound*: fear or elation without anything presented as fearful or good. That makes her an ally against contentlessness and a rival to the body-and-field reading I favor. Her contrast has force. Feeling sad need not feel like finding everything sad. A body-and-field account must explain that difference rather than make every mood a verdict on the whole world. The two-layer distinction in Section 6.5 gives us a constraint: a particular object fixes singular reference and accuracy; character is the presented property profile under its mode, and on my wide view history helps fix which properties that profile contains. It does not settle what affective properties moods present, or whether a mood binds them to anything. Mendelovici's account therefore remains a live alternative.[10] Neither naming an unbound property nor naming the whole world supplies the independent account of affective content this book needs.

Try to attend to your bad mood directly: look *at* it rather than *through* it, and catch the dread apart from the dingy kitchen, the irritating cat, and the day-to-be-gotten-through. Trained attention may find tightness, heaviness, agitation, or a diffuse affective tone. What it does not uncover is a detachable inner object. It finds more body, more world, or the affective mode under which both appear. The mood need not slip every time for transparency to matter. It need only refuse the role of mental paint displayed behind the scene.

But a serious rival can grant every change in the kitchen and deny that the mood itself represents anything. A nonintentional affective background might *cause* the coffee to seem tasteless, the conversation remote, and the future closed. It need not resemble a patch on an inner screen, and it need not lack structure. Showing that mood changes content does not show that mood consists in content.

I favor the situated presentation because it joins the bodily weight and the changed possibilities in one way of meeting the day. But consider the kitchen again. A mood might present the day as bleak, or it might cause the coffee to seem tasteless and conversation to feel remote without itself presenting bleakness. The altered scene alone cannot choose between those accounts. We need an independent account of what the mood gets right or wrong. Chapter 15 offers candidates, not that completed account. The kitchen supports a line of inquiry; it does not add a solved case to the perceptual argument for identity.

**Chapter Summary**

**The claim.** I defend the identity claim for pain and mood as well as perception. Their apparent lack of an object does not refute it, but neither has this chapter established what their affective content consists in.

**Where the argument stands.** The bodily-address account explains how phantom pain can get location wrong. Asymbolia does not decide between evaluative content and aversive reaction; the kitchen scene does not decide between a mood with content and a feeling that changes other contents. The book's stronger explanatory support remains with its perceptual cases.

**What's left open.** Pain's badness, mood's content, and the grounds for ending an experience itself need accounts that do not borrow their answer from how those states feel. Chapter 15 supplies tests and candidates rather than discharging those debts. Chapter 10′s comparison with a further phenomenal ingredient cannot discharge them either.

**The hand-off.** Pain and mood leave the universal thesis under pressure, not refuted by their lack of an obvious object. Mary in her black-and-white room poses a different challenge: what can experience teach us?

**Footnotes**

**1.** V. S. Ramachandran and Sandra Blakeslee, *Phantoms in the Brain: Probing the Mysteries of the Human Mind* (New York: William Morrow, 1998), chs. 1–3, on phantom limbs, the remapping of body representation in somatosensory cortex, and the mirror-box relief of phantom cramp. The prevalence claim rests on survey evidence: in a national survey of people with limb loss, 79.9 percent reported phantom pain in the previous four weeks (Patti L. Ephraim et al., "Phantom Pain, Residual Limb Pain, and Back Pain in Amputees: Results of a National Survey," *Archives of Physical Medicine and Rehabilitation* 86, no. 10 [2005]: 1910–1919, doi:10.1016/j.apmr.2005.03.031). See also Ronald Melzack, "Phantom Limbs and the Concept of a Neuromatrix," *Trends in Neurosciences* 13, no. 3 (1990): 88–92, whose neuromatrix model independently supports the chapter's point that the felt body is a construction the brain maintains with or without the limb — which is why the representation can persist, and misfire, after amputation.

**2.** The representational account of pain followed here is developed in Michael Tye, *Ten Problems of Consciousness* (Cambridge, MA: MIT Press, 1995), ch. 4, and "Another Look at Representationalism about Pain," in *Pain: New Essays on Its Nature and the Methodology of Its Study*, ed. Murat Aydede (Cambridge, MA: MIT Press, 2005), 99–120. Tye's PANIC framework treats pain as poised, abstract, nonconceptual, intentional content directed at a bodily region; the chapter adopts the structural claim while stating its own provisional content more carefully as an actual or impending threat

to bodily integrity at a represented location, presented as bad for the subject under a bodily-affective mode.

**3.** The parallel between phantom pain and misrepresentation in vision (the bent stick, the mirage) connects to the error machinery the book grounds in Part Three. What makes a pain state present *an actual or impending threat to bodily integrity at the toe* — rather than merely arise from toe damage — is the teleofunctional account of content developed in Chapter 15 (Millikan, Dretske): the state has the proper function of indicating a threat at a bodily location, which allows it to misrepresent when it is produced without the corresponding condition. The phantom limb is to the body-monitoring system what the frog's misfiring fly-detector is to the frog: a representation produced in the absence of the condition it has the function to indicate. Where stump or nerve pathology drives the phantom, the case resembles referred pain instead: a real disturbance assigned the wrong place. Either way the presented location is wrong. Whether such content concerns a threat at the toe, tissue damage, or nerve activity is the grain problem the frog made famous; Section 15.3 addresses it for the frog, and the bodily case inherits it. Section 15.3.2 develops this proposal for bodily address while distinguishing it from the unfinished naturalization of evaluative badness, mood, and experience-directed relief. The address proposal does not discharge those further debts.

**4.** David Bain, "Intentionalism and Pain," *Philosophical Quarterly* 53, no. 213 (2003): 502–523, and "What Makes Pains Unpleasant?," *Philosophical Studies* 166, suppl. 1 (2013): S69–S89, develop the evaluativist position: the unpleasantness of pain consists in the experience's representing the bodily disturbance as *bad for* the subject, where the badness is representational content, not a non-intentional residue. Bain's response to the standard objection (that a pain could represent damage without feeling bad) is that the felt badness just *is* a further layer of (evaluative) content. Section 7.4 relies on that move, placing the evaluative content under a bodily-affective mode. Cutter and Tye also give pain evaluative content but make it trackable: pain represents a disturbance as bad for the subject to some degree, where being bad for the subject is identified with being apt to harm her. See Brian Cutter and Michael Tye, "Tracking Representationalism and the Painfulness of Pain," *Philosophical Issues* 21 (2011): 90–109, sec. 3.

**5.** Colin Klein, *What the Body Commands: The Imperative Theory of Pain* (Cambridge, MA: MIT Press, 2015), and "An Imperative Theory of Pain," *Journal of Philosophy* 104, no. 10 (2007): 517–532; see also Manolo Martínez, "Imperative Content and the Painfulness of Pain," *Phenomenology and the Cognitive Sciences* 10, no. 1 (2011): 67–90. Imperative content has a world-to-mind direction of fit: it commands rather than describes. That differs directly from the truth-apt locative and evaluative content the chapter needs for ordinary and phantom pain, so the house view treats imperative force as a friendly rival or as a candidate mode, a manner of presenting rather than further content, not an interchangeable formulation. Hilla Jacobson, "Killing the Messenger: Representationalism and the Painfulness of Pain," *Philosophical Quarterly* 63 (2013): 509–519, poses the action-directed challenge. Brian Cutter and Michael Tye answer it in "Pains and Reasons: Why It Is Rational to Kill the Messenger," *Philosophical Quarterly* 64, no. 256 (2014): 423–433, doi:10.1093/pq/pqu025. Matthew Kinakin sharpens both the motivational and epistemic problems in "Why Intentionalists Can't Take Painkillers," *Philosophical Studies* 183 (2026): 1905–1929, doi:10.1007/

s11098-026-02523-z. The text considers a first-order gain-control answer but acknowledges its limits: pain may warrant relief even without further impairment, and a second-order evaluation's having content does not explain its epistemic warrant. Neither point by itself establishes nonrepresentational phenomenal paint. The positive account of that warrant remains unfinished.

**6.** The classic philosophical treatment is Nikola Grahek, *Feeling Pain and Being in Pain*, 2nd ed. (Cambridge, MA: MIT Press, 2007), which assembles asymbolia and related "reactive dissociation" cases. For the lesion evidence, see Marcelo Berthier, Sergio Starkstein, and Ramón Leiguarda, "Asymbolia for Pain: A Sensory-Limbic Disconnection Syndrome," *Annals of Neurology* 24, no. 1 (1988): 41–49, doi:10.1002/ana.410240109, which found insular damage in all six patients described. The classification now stands contested. Trevor Griffith and Adrian Kind, "Pain Asymbolia Is Not Pain," *Philosophy of Science* 91, no. 3 (2024): 561–578, doi:10.1017/psa.2023.167, argue that the clinical evidence does not support pain-without-painfulness. Alexandre Duval and Colin Klein, "Pain Asymbolia Is Probably Still Pain," *Philosophy of Science* 93, no. 1 (2026): 221–229, doi:10.1017/psa.2025.10098, defend the standard taxonomy while emphasizing what remains unresolved. The body therefore uses the dissociation comparatively and does not treat the syndrome's name as a settled description of its phenomenology.

**7.** Murat Aydede, "Is Feeling Pain the Perception of Something?," *Journal of Philosophy* 106, no. 10 (2009): 531–567, doi:10.5840/jphil20091061033, presses the full-strength challenge to perceptual and strong representational theories of pain. For the scientific and conceptual background, see Murat Aydede and Güven Güzeldere, "Some Foundational Problems in the Scientific Study of Pain," *Philosophy of Science* 69, no. 3S (2002): S265–S283. A kindred resistance, on which pain includes nonrepresentational "mental paint," appears in Ned Block, "Mental Paint," in *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, ed. Martin Hahn and Bjørn Ramberg (Cambridge, MA: MIT Press, 2003), 165–200. The phantom-bruise comparison belongs to this book and functions as abductive evidence only; Chapter 15 owes the proper-function account that warrants accuracy talk.

**8.** David Bourget and Angela Mendelovici, "Tracking Representationalism: Lycan, Dretske, and Tye," in Andrew Bailey, ed., *Philosophy of Mind: The Key Thinkers* (London: Bloomsbury, 2014), identify "elation, free-floating anxiety, or pervasive sadness" as cases that do not readily qualify objects. William G. Lycan argues instead, in *Consciousness and Experience* (Cambridge, MA: MIT Press, 1996), that even the vaguest moods carry definite, if comparatively uninteresting, representational content; he restates the claim in "The Intentionality of Smell," *Frontiers in Psychology* 5 (2014): 436, n. 6, doi:10.3389/fpsyg.2014.00436. Angela Mendelovici, "Intentionalism about Moods," *Thought* 2, no. 2 (2013): 126–136, doi:10.1002/tht3.81, states the apparent counterexample at full strength before answering it.

**9.** Matthew Ratcliffe, *Feelings of Being: Phenomenology, Psychiatry and the Sense of Reality* (Oxford: Oxford University Press, 2008), develops "existential feelings" as background bodily-affective orientations that shape the sense of reality and belonging. When disturbed, they show up in depression and depersonalization as a changed *way the world is given* rather than a discrete inner object. Ratcliffe does not endorse Tyean representationalism; the chapter cites him as a convergent witness on the phenomenology, not an ally on the metaphysics.

**10.** Michael Tye, *Ten Problems of Consciousness* (Cambridge, MA: MIT Press, 1995), esp. sec. 4.10, and *Consciousness, Color, and Content* (Cambridge, MA: MIT Press, 2000), esp. 51, reads moods as representations of global states of the body or situation. Mendelovici, "Intentionalism about Moods," cited n. 8, develops the unbound-affective-properties account: some moods represent affective properties such as threat or elation without representing an object as bearing them. That view supports intentional content while resisting this chapter's preferred object-involving interpretation. Mendelovici defends it as *pure* intentionalism, on which phenomenal character is exhausted by what is represented: "Pure Intentionalism about Moods and Emotions," in *Current Controversies in Philosophy of Mind*, ed. Uriah Kriegel (New York: Routledge, 2014), 135–157. The house view is impure, since it counts the bodily-affective mode in the profile (Section 7.4); her account is therefore a rival on that point as well as on binding. Her later *The Phenomenal Basis of Intentionality* (New York: Oxford University Press, 2018) reverses this book's order of explanation, giving phenomenology priority over intentionality.

## Chapter 8: The Knowledge Argument

> *"It seems just obvious that she will learn something about the world and our visual experience of it."*
>
> — Frank Jackson, "Epiphenomenal Qualia"

CHAPTER OVERVIEW

Mary knows every physical fact about color in a black-and-white room, yet the first red tomato still teaches her something. I grant that result, since denying it saves physicalism by throwing away the pull of the case. The key step comes next: new knowledge need not add a new kind of fact to the world. New skills, direct acquaintance, and a first-person route can bring one physical fact into view in a new way. Mary can now spot, picture, and recall red, and she can form a true thought she could not form before; neither gain shows that physics left an extra property off its list. My answer keeps both her learning and one physical world, while owing a fuller account of why no report can give the knowledge gained through seeing.

Here is a room without color: gray walls, a black-and-white monitor, a black-and-white book. Mary has never seen a color, but she has mastered the completed physics, neuroscience, and science of color vision. She knows the wavelengths, the cones and ganglion cells, every cortical event, and every word an English speaker reaches for after looking at a ripe tomato. She knows every physical fact about seeing red, all in grayscale.

Then one day the door opens, and Mary walks out, and someone hands her a tomato.

Now: does Mary learn anything?

Almost everyone says yes. It seems she learns *what red looks like* — and that she could not have learned it inside, no matter how thick her textbooks, because the room held every physical fact and this was not in them. And if Mary knew all the physical facts and still had something to learn, then the physical facts are not all the facts. Something about seeing red, the very thing she now has and lacked before, escapes the complete physical story. Consciousness, it seems, is the one corner of nature that a finished physics would leave out.

That is the Knowledge Argument, and it has earned the respect of the very side it attacks. Mary leaves the room changed. What she does not show, though it takes some care to see why, is that the world contained a non-physical fact all along.

### 8.1 The Argument at Full Strength

We owe the thought experiment to Frank Jackson, an Australian philosopher who would later reject the argument that made Mary famous.[1] He introduced her in a 1982 paper with the blunt title "Epiphenomenal Qualia," and sharpened the case four years later in "What Mary Didn't Know."[2] Philosophers have refined and attacked it for forty years, and it still looms over physicalist theorizing about the mind.[3]

Set it out plainly. Jackson asks us to grant three things:

1. Mary, before she leaves, knows *all the physical facts* about color vision — every fact statable in the completed vocabulary of physics, chemistry, and neuroscience. The stipulation buys an ideal knower, not a guess about real neuroscientists, so that nothing physical stays in reserve.
2. On leaving the room and seeing red for the first time, Mary *learns something* — she comes to know what it is like to see red, which she did not know before.
3. Here is the hinge: what she learns is a *fact*. A truth about the world she did not previously possess. She used to be ignorant of it; now she knows it.

From these three the conclusion drops out almost on its own.

> If Mary knew all the physical facts (premise one) and still learned a fact on her release (premises two and three), then the fact she learned was not among the physical facts. So there is at least one non-physical fact. So physicalism about experience (the claim that experience adds no irreducibly non-physical feature to reality) is false. The lights of consciousness are real, and they fall outside the circle physics can draw.

The argument needs no ghost or immaterial substance: just a scientist, a tomato, and the thought that she learns something. It seems to read dualism straight off an ordinary fact about learning. You need not want the conclusion to feel the floor tilt.

Daniel Dennett asks us not to grant even that ordinary fact too quickly. A gifted American philosopher with a talent for exposing what a thought experiment quietly asks us to imagine, he argues that we replace Jackson's ideal knower with a merely brilliant neuroscientist and then trust the surprise we have built into the substitute. Mary knows *all* the physical information, including every effect color will have on her. RoboMary sharpens the point: perhaps an ideal artificial scientist could use that information to reconfigure herself and identify red in advance.[4] The challenge puts premise two, not premise three, on trial. The argument here grants that Mary learns something. Complete propositional knowledge does not obviously install every representational vehicle or cognitive relation; if RoboMary induces the relevant color state in herself, she has found an unusual route to undergoing it. Still, Dennett has earned a caution: the familiar gasp of surprise is not sacrosanct.

This chapter grants Mary the stronger case. She gains something. The question is whether that gain forces a new item into the world's inventory. A hundred tidier objections have come and gone. Mary still stands.

### 8.2 The Author Against the Argument

Then the man who built Mary tore her down.

By the late 1990s Frank Jackson had stopped believing his own argument. He became a physicalist and came to deny that Mary acquires a new truth about how the world is. His mature answer ran through strong representationalism: release puts Mary into a new kind of representational state

and gives her new capacities, without revealing a further property outside the physical story.[5] That answer is more deflationary than the one this chapter will defend. I will grant Mary a genuinely new cognitive achievement and ask what follows from it.

Jackson published the reversal more than once and developed the positive account that led him there.[6]

His reversal does not settle the argument. Its interest lies in the demand Jackson now tried to meet: explain why release changes Mary so profoundly without treating that change as contact with a second order of reality. One pressure came from epiphenomenalism's strained relation to our reports. On that view, the non-physical quality accompanies the neural activity without influencing it. Yet Mary walks into the garden and says, *so this is what red looks like.* The old view threatens to make the redness of her world causally irrelevant to her report of it. Whatever the exact chronology of Jackson's turn, that self-stultifying result gave him reason to seek a physical home for the datum; representationalism supplied his published route.

The route here parts company with Jackson at one point. It grants that Mary learns something true and asks whether that truth requires a new non-physical truthmaker. This more concessive type-B strategy keeps the epistemic gap without letting it cross into ontology. The anthologies, meanwhile, keep Jackson's name on the argument.

### 8.3 The Missing Bridge: From Epistemic to Ontological

Mary learns what seeing red is like. She can think, *so this is the kind of experience other people have been having when they look at ripe tomatoes.* The knowledge reaches beyond the tomato in her hand and beyond this morning: it concerns a kind of experience she can recognize again and attribute to others. I grant that she learns something true, not merely that she gains a knack.

The disputed step concerns what makes that new thought true. The Knowledge Argument needs this bridge:

> *If someone gains propositional knowledge unavailable under a complete physical description, that knowledge must be made true by some non-physical feature of reality.*

Only that bridge secures the conclusion. Jackson knew the danger. Paul Churchland, who spent a career arguing that some of our everyday mental vocabulary may eventually look the way talk of phlogiston looks now, charged that the argument slides between knowledge by description and knowledge by acquaintance. Jackson answered in 1986 by saying that Mary gains information, knowledge of a fact, precisely to block the suggestion that she merely acquires a new skill or acquaintance.[7] A mature reply should grant that formulation. Calling *fact* ambiguous will not do the work by itself.

Mary's true thought expresses a *proposition*, something she can believe or know. Other people's seeing red supplies a *state of affairs*, how things stand in the world. Whatever makes her thought true supplies its *truthmaker*. These terms separate the new thought from what it concerns. Two

thoughts can reach one state of affairs; acquiring the second need not reveal an extra property. Whether Mary's thoughts do that remains the dispute.

Complete physical knowledge, as stipulated here, leaves no physical condition or relation undescribed. It does not automatically give Mary every way of thinking about those conditions. The dualist objects that this qualification retreats from *all the facts*: if seeing teaches a truth, her earlier knowledge was incomplete. The physicalist agrees that her knowledge changes but denies that completeness of the world's physical basis requires completeness of every knower's routes to it. That disagreement needs argument. Neither side can settle it by repeating *all*.

The city map makes the logical point. Suppose you know every street, building, and lamppost from the most detailed map ever drawn, then someone sets you down in the central square. You can now think *I am here*, placing yourself among streets you already knew. No new street sprang up when you arrived. The map drew every street; it could not stand you in the square.

That analogy does not prove that Mary's physical map exhausts reality. The dualist denies exactly that. It shows something narrower and indispensable: a genuine epistemic change need not add a worldly item. The move from *Mary learns* to *the physical inventory omitted something* therefore cannot be automatic. The argument still owes the bridge.

Chapter 9 meets the same architecture with different starting material. Mary's gap concerns modes of access. The zombie argument grants, as this book does, that a physical duplicate without experience remains ideally conceivable, then seeks a modal conclusion: such a duplicate is metaphysically possible. Neither case reduces to a temporary failure of imagination, and neither uses the same vocabulary. They share the need for a further principle carrying an epistemic datum into ontology.[8]

### 8.4 What Mary Gains

The first clean answer came from David Lewis, in a paper called "What Experience Teaches" that first appeared not in a flagship journal but in the proceedings of a philosophy society at the University of Sydney.[9] Lewis had a habit of taking a wild-sounding thesis and arguing it, by the end of the paper, into the only sober option on the table; here he ran the trick in reverse and made a sober reply look like the obvious one. The core suggestion came from Laurence Nemirow, who first floated it in a review of Thomas Nagel's *Mortal Questions*: Mary gains not knowledge *that* anything is the case but knowledge *how* — a bundle of abilities.[10]

Before release Mary cannot recognize red by sight, summon a mental image of red, or remember red experiences as such. Afterward she can. Experience teaches skills: to recognize, imagine, and remember. You can know a bicycle's mechanics before you can ride one; the first successful lap adds no non-physical property to the bicycle. Acquiring Mary's new skills likewise need not add such a property to what she encounters.

The ability hypothesis breaks premise three's spell by displaying a familiar kind of learning that does not threaten physicalism. It also identifies a real component of Mary's change. Whatever else release gives her, it equips her to do things she could not do inside.

That component does not exhaust the case. Mary seems to *grasp something*, not merely acquire a knack. Even philosophers friendly to physicalism have resisted the reduction to know-how. Michael Tye, whose representationalism this book otherwise draws on, argues that Mary's new state puts her in cognitive contact with phenomenal character in a way "she can now imagine it" underdescribes.[11] Earl Conee gave the acquaintance reply one of its clearest formulations: acquaintance itself supplies the epistemic gain. Those views mark a live debt. Calling acquaintance the route by which a new concept forms does not yet show that acquaintance contributes no distinctive knowledge.

The ability hypothesis therefore belongs inside the eventual answer. It describes Mary's change at the level of skill; whether her apparent propositional gain can remain neutral about the world's inventory requires a further account.

### 8.5 A Model: One Fact, Two Modes of Access

A physicalist proposal runs this way: Mary comes to know an old worldly fact under a new mode of presentation.

Saul Kripke's 1970 Princeton lectures made necessary identities discovered through experience central to this debate. The familiar example, *water is $H_2O$*, supplies the model: an informative discovery without a second substance.[12] But Mary already knows the completed physical account. Repeating a scientific discovery story cannot establish that her new phenomenal thought concerns that same physical reality.

The physicalist model says that Mary knew color experience theoretically, through wavelengths, cones, cortex, and behavior. On release she encounters the same state through a first-person route: the way seeing red presents itself to someone undergoing it. Her cry of recognition, *so this is what they were talking about*, may record two roads meeting at one place. The ability account describes the new recognitional equipment. The modes account describes a possible content for the recognition. In its compressed form: the knowledge was new. The known was not.

That remains a proposal. The physicalist has shown that the argument's conclusion is not forced; she has not yet shown that this proposal supplies the best account of Mary. Knowing every physical fact does not guarantee standing in every cognitive relation to those facts, but the dualist can still argue that the new relation discloses something the physical facts left out. The burden has moved from the sheer existence of learning to the nature and truthmaker of what is learned.

Critics of the old-fact/new-mode view often put the worry this way: calling the two modes co-referential assumes what the dualist denies. Martine Nida-Rümelin has kept the pressure on the point where experience seems to disclose its own nature. Her stronger claim says that phenomenal concepts do not merely point at their properties; they let a thinker *grasp what those properties essentially consist in*. Her principle of cognitive transparency then presses the physicalist: if Mary's new concept reveals a property's nature, the relation between that property and the complete physical story should not remain opaque in the way water's relation to $H_2O$ once did.[13] The physicalist cannot answer that challenge by relabeling opacity as a mode of presentation.

Return to the city square. *I am here* gives the visitor new knowledge tied to a standpoint, what philosophers call *indexical* knowledge. Mary's *this* likewise fastens on the experience she is undergoing. John Perry develops that route to a physicalist reply. Its limit now matters: a map coordinate can settle the visitor's location, whereas Mary's central deficit appears to close only through experience. And she learns about others' seeing, not just her own present position. Chalmers presses both differences against Perry. Mary's demonstrative may carry a qualitative concept that reveals a nature rather than merely locating her within an order.

The city therefore illustrates the logical room for a reply, not a complete reply. Chapter 11 must explain why undergoing matters, how Mary's knowledge generalizes, and why grasping phenomenal character need not disclose a non-physical essence.

### 8.6 What the Resolution Defers — and the Bridge Forward

Jackson's route from Mary's epistemic change to a non-physical fact needs a bridge the learning premises alone do not establish. He argues for that bridge; the physicalist disputes it. Ability, a-posteriori identity, and indexical models show why the inference cannot be automatic without denying Mary's change.

The positive account remains conditional: Mary meets one physical reality through a new first-person route. A completed explanation must say why undergoing a state makes a concept available when complete description does not, why the resulting thought can generalize beyond the present subject, and why phenomenal grasp does not reveal a non-physical essence. The indexical analogy names a logical possibility; it does not pay those bills.

The *phenomenal-concept strategy* offers one way to pay them: human beings possess concepts formed through experience that can refer to physical states by a route our theoretical concepts do not travel.[14] The narrower house proposal needs less. Acquaintance with the occurring state grounds recognitional capacities — seeing red again, imagining it, remembering it as such — without requiring a proprietary class of concepts to carry the entire explanation. That proposal still needs defense. Tye's mature materialism rejects special phenomenal concepts while retaining a role for acquaintance, Conee treats acquaintance as an epistemic category in its own right, Nida-Rümelin argues that phenomenal grasp reveals a property's nature, and Chalmers denies that an indexical component captures Mary's substantive new knowledge.

Chapter 11 takes up the acquaintance-recognition account and the surrounding dispute in full.[15] Its task is to explain why descriptive and first-person routes feel as far apart as they do, and why that distance may mark a fact about our access rather than a fault line in the world. The Knowledge Argument has not established ontological novelty; the physicalist account of Mary's epistemic novelty still has work to do.

### Chapter Summary

**The claim.** Mary's release gives us a genuine fact about knowledge, but the Knowledge Argument does not turn it into a non-physical fact by itself. That conclusion needs a bridge from new knowledge to a new truthmaker in the world.

**Where the argument stands.** The ability hypothesis captures a real part of Mary's change. Cases of a posteriori and indexical knowledge also show that thought can change without adding anything to reality. These cases make a physicalist account possible; they do not yet show that Mary's new knowledge fits one.

**What's left open.** Chapter 11 must explain why undergoing the state matters. It must answer Nida-Rümelin's grasping challenge and Chalmers's objection to the indexical analogy, while capturing how Mary's knowledge reaches beyond her own case. It must also explain what acquaintance adds without making any one theorist's account of phenomenal concepts the house view.

**The hand-off.** The next chapter moves from something *left out* to something *subtracted*: the conceivable zombie, your physical duplicate with the lights off inside. It grants ideal conceivability and asks what permits the move to metaphysical possibility. Bring a flashlight.

**Footnotes**

**1.** Frank Jackson (b. 1943), for years professor at the Australian National University (which he joined in 1986), made his name as an uncompromising defender of following arguments to their conclusions; his later reversal on the Knowledge Argument is itself a case study in that disposition turned reflexive. The epigraph is from Jackson, "Epiphenomenal Qualia," *The Philosophical Quarterly* 32, no. 127 (1982), 130.

**2.** Frank Jackson, "Epiphenomenal Qualia," *The Philosophical Quarterly* 32, no. 127 (1982): 127–136, esp. 130; and "What Mary Didn't Know," *The Journal of Philosophy* 83, no. 5 (1986): 291–295, esp. 291–94. The 1982 paper introduces Mary; the 1986 paper restates the argument as explicitly about knowledge of facts and answers early replies.

**3.** Central early replies and counter-replies are collected in Peter Ludlow, Yujin Nagasawa, and Daniel Stoljar, eds., *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument* (Cambridge, MA: MIT Press, 2004). For the current reference treatment, see Martine Nida-Rümelin and Donnchadh O Conaill, "Qualia: The Knowledge Argument," *Stanford Encyclopedia of Philosophy* (substantive revision March 1, 2024).

**4.** Daniel C. Dennett, *Consciousness Explained* (Boston: Little, Brown, 1991), 398–414, argues that the standard case under-imagines the consequences of Mary's complete physical information; he offers the blue-banana variant as a counter-story rather than a proof that Mary learns nothing. "What RoboMary Knows," in *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, ed. Torin Alter and Sven Walter (New York: Oxford University Press, 2007), 15–31, develops the self-reconfiguration challenge, DOI 10.1093/acprof:oso/9780195171655.003.0001.

**5.** Jackson's mature position denies that release gives Mary new propositional knowledge about how things are. His positive route uses strong representationalism and a Lewis/Nemirow-style account of the capacities supplied by a new representational state. See Jackson, "Mind and Illusion" (cited in n. 6 below), esp. "Back to Mary on her release." The present chapter's willingness

to grant Mary a genuinely new true thought belongs to this book's more concessive type-B reconstruction, not to Jackson's mature view.

**6.** Jackson's recantation is developed across several pieces, notably "Postscript on Qualia," in *Mind, Method, and Conditionals: Selected Papers* (London: Routledge, 1998), 76–79, and "Mind and Illusion," in *Minds and Persons*, ed. Anthony O'Hear, *Royal Institute of Philosophy Supplement* 53 (Cambridge: Cambridge University Press, 2003), 251–272, esp. 251–52 and the section "Back to Mary on her release." The causal impotence of epiphenomenal qualia supplies a genuine pressure on the earlier view; the text does not claim it records the sole historical cause of Jackson's reversal.

**7.** Paul M. Churchland, "Reduction, Qualia, and the Direct Introspection of Brain States," *The Journal of Philosophy* 82, no. 1 (1985): 8–28, esp. 23–24, charges that the Knowledge Argument equivocates between knowledge by description and knowledge by acquaintance. Jackson's "What Mary Didn't Know," esp. 293–94, answers by casting the argument in terms of factual information. Churchland renews the dispute in "Knowing Qualia: A Reply to Jackson," in *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science* (Cambridge, MA: MIT Press, 1989), 67–76. The text therefore grants the propositional formulation and locates the disputed step in the bridge from a new proposition to a non-physical truthmaker.

**8.** Chapter 9, *Why Dualism Keeps Winning Arguments It Should Lose*, grants the type-B datum that zombies remain ideally conceivable and questions the further inference to metaphysical possibility. The parallel with Mary concerns argumentative architecture, not identical premises or a merely temporary limit on imagination.

**9.** David Lewis (1941–2001), among the most systematically inventive analytic philosophers of his century, was known for defending startling theses — the reality of possible worlds among them — with such rigor that the startle wore off; his reply to Jackson is unusual in his corpus for making the deflationary option the compelling one.

**10.** David Lewis, "What Experience Teaches," originally in *Proceedings of the Russellian Society* 13 (1988): 29–57, reprinted in *Mind and Cognition: A Reader*, ed. William G. Lycan (Oxford: Blackwell, 1990), 499–519, esp. "The Ability Hypothesis." Lewis credits Laurence Nemirow, whose earlier statement appears in his review of Thomas Nagel's *Mortal Questions*, *The Philosophical Review* 89, no. 3 (1980): 473–477; see also Nemirow, "Physicalism and the Cognitive Role of Acquaintance," in Lycan, ed., *Mind and Cognition*, 490–499.

**11.** Michael Tye rejects the bare ability hypothesis as a complete account of Mary's gain in *Consciousness Revisited: Materialism Without Phenomenal Concepts* (Cambridge, MA: MIT Press, 2009), esp. 125–33. That book also rejects the standard phenomenal-concept strategy, esp. 56, while assigning acquaintance a materialist explanatory role. The present book agrees that pure know-how is insufficient and takes acquaintance seriously without treating Tye's full account as authoritative. See Tim Crane, "Tye on Acquaintance and the Problem of Consciousness," *Philosophy and Phenomenological Research* 84, no. 1 (2012): 190–198, esp. 190–96. Earl Conee, "Phenomenal Knowledge," *Australasian Journal of Philosophy* 72, no. 2 (1994): 136–150, esp. 144, treats acquaintance as a distinctive epistemic relation; Chapter 11 must address rather than redescribe that claim.

**12.** The two-modes model descends from Saul Kripke's treatment of necessary a-posteriori identities in *Naming and Necessity* (Cambridge, MA: Harvard University Press, 1980), esp. 116–17 and 128–29. The water/$H_2O$ case establishes that informative identity is possible; it does not establish the psycho-physical identity at issue in Mary's case.

**13.** Martine Nida-Rümelin's distinctive challenge concerns phenomenal *grasp*: phenomenal concepts allegedly reveal what possessing their referents essentially consists in, and her cognitive-transparency principle pressures opaque physical/phenomenal identities. See "Grasping Phenomenal Properties," in *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, ed. Torin Alter and Sven Walter (New York: Oxford University Press, 2007), 307–338, esp. secs. 1, 13, and 15; and Nida-Rümelin and Donnchadh O Conaill, "Qualia: The Knowledge Argument," *Stanford Encyclopedia of Philosophy* (substantive revision March 1, 2024), under "The New Knowledge/Old Fact View." For the indexical model, see John Perry, "The Problem of the Essential Indexical," *Noûs* 13, no. 1 (1979): 3–21, and *Knowledge, Possibility, and Consciousness* (Cambridge, MA: MIT Press, 2001). David Chalmers, "Imagination, Indexicality, and Intensions," *Philosophy and Phenomenological Research* 68, no. 1 (2004): 182–190, esp. sec. 2, presses the experience-dependence and generality objections, DOI 10.1111/j.1933-1592.2004.tb00334.x.

**14.** The phenomenal-concept strategy holds that phenomenal concepts can be cognitively isolated from physical concepts while co-referring with them. See Brian Loar, "Phenomenal States," *Philosophical Perspectives* 4 (1990): 81–108, revised in *The Nature of Consciousness: Philosophical Debates*, ed. Ned Block, Owen Flanagan, and Güven Güzeldere (Cambridge, MA: MIT Press, 1997), 597–616, esp. secs. 2–3; David Chalmers, "Phenomenal Concepts and the Explanatory Gap," in Alter and Walter, eds., *Phenomenal Concepts and Phenomenal Knowledge* (New York: Oxford University Press, 2007), 167–194, esp. secs. 2–3; Robert J. Howell, "The Knowledge Argument and the Implications of Phenomenal Knowledge," *Philosophy Compass* 6, no. 7 (2011): 459–468, DOI 10.1111/j.1747-9991.2011.00408.x; and Howell, *Consciousness and the Limits of Objectivity: The Case for Subjective Physicalism* (Oxford: Oxford University Press, 2013), DOI 10.1093/acprof:oso/9780199654666.001.0001. Michael Tye rejects this strategy in *Consciousness Revisited*, esp. 56; that disagreement remains live.

**15.** Chapter 11, *The Gap That Isn't There*, develops the house acquaintance-recognition account, locates it within the phenomenal-concept debate, registers Tye's rejection of special phenomenal concepts, and applies the two-routes/one-fact structure to the explanatory gap as a whole. Chapter 8 supplies only the negative result and the shape of a possible positive account. Sections 11.3 and 11.7 limit the acquaintance-recognition explanation to concept acquisition and possible conceptual asymmetry; successful reference, justification, and the whole truth-involving epistemic gap remain further tasks. The identity's independent warrant cannot be supplied by the acquisition account.

## Chapter 9: Why Dualism Keeps Winning Arguments It Should Lose

> *"There is nothing that we know more intimately than conscious experience, but there is nothing that is harder to explain."*
>
> — David Chalmers, "Facing Up to the Problem of Consciousness"

CHAPTER OVERVIEW

Dualism starts from truths a physicalist should keep: experience has a point of view, and no account from outside can give us the bat's own view. I grant that even after ideal thought we can conceive of a full physical twin with no felt life, but I deny that this twin could exist. What thought tells us about what could exist remains the point at issue, not our power to form the thought. A second test asks what a distinct mental cause could do if physical causes already suffice: change their course, cause the same effect again, or do nothing. Causing an effect twice differs from doing nothing, so each choice needs its own reply. I argue that felt life belongs to our physical life and its world-shaped past; a copy of present inner roles alone leaves that past out.

Consider the bat.

Thomas Nagel asks us to imagine what it would be like to be a bat—navigating by echolocation, building a sonic picture of the world from returning cries.[1] We can know its nervous system, acoustic processing, and spatial maps in exquisite detail. We can even imagine hanging upside down in a dark cave. What we cannot imagine, Nagel says, is what it *feels like* from the bat's point of view to inhabit that moving, three-dimensional sound-map. There is a fact about what that experience is like. We cannot access it from outside.

That is the point Nagel wants to make, and it cuts deep. The problem isn't that we lack data about bats. It's that all the data in the world (every neuron mapped, every acoustic pathway charted, every behavior explained) describes the bat from a third-person perspective, as an object in the world. But experience has a first-person character. It presents itself from somewhere, to someone. And that somewhere, that to-whom, seems to fall clean outside anything science's objective vocabulary can reach.

Nagel isn't trafficking in mysticism. He isn't claiming that bats have immaterial minds, or that experience floats free of the physical world in some ghostly parallel realm. His claim is more modest and, for that reason, more difficult to dismiss: there is something it is like to be a bat, and no third-person description fully captures what that something is. The subjective character of experience, the *what-it's-likeness*, seems to resist reduction to the terms physics and neuroscience use to describe everything else.

Most people, on reflection, find this compelling, and with reason. The mistake comes later, in what we ask the observation to explain.

### 9.1 The Hard Problem and the Zombie

Chapter 2 stated the Hard Problem whole. Here is its sharpest formal edge: Chalmers's zombie — a physical duplicate of a human being, molecule for molecule identical, with the lights off inside. All the neural processing happens, all the behavior results, and nothing it is like to be it. The argument from the zombie runs clean.

> If such a creature is conceivable, Chalmers holds, it is possible; and if a physical duplicate could lack experience, then experience must be something over and above the physical. Dualism of some kind follows.

The conclusion still depends on the bridge from conceivability to possibility.

The chat window gives the question a practical setting. A fluent reply leaves open what, if anything, the system feels. That uncertainty does not by itself support either verdict. Part Four asks which facts about an actual system could carry the inquiry beyond what we can imagine about it.

### 9.2 The Diagnosis: What Makes the Gap Feel Unbridgeable

Joseph Levine coined *explanatory gap* for a problem distinct from Chalmers's Hard Problem.[2] Grant that consciousness involves nothing beyond the physical and that pain amounts to a neural process—C-fiber stimulation in the textbook caricature, or some more complex state neuroscience discovers. We can explain why that process produces avoidance, recruits attention, and integrates with memory, yet still seem unable to explain why it *hurts*. A complete physical story leaves felt character hanging. Levine's gap lies in our understanding rather than the world, and troubles even committed physicalists.

Chalmers gave Levine's gap an ontological consequence, but he did not get there by stamping *metaphysical* on an epistemic problem. The zombie argument supplies the bridge. His premise concerns *ideal primary positive conceivability*: a coherently imagined scenario that survives ideal rational reflection and verifies the full physical truth without consciousness when the scenario gets considered as actual. Chalmers argues that such conceivability entails primary possibility; if a possible world verifies the scenario, the physical facts do not fix all the facts.[3]

One especially dramatic rendering of the Hard Problem feels unbridgeable because it misdescribes both sides. On one side stand physical processes, described in entirely third-person, functional, structural vocabulary. On the other side stands phenomenal experience, described as inner, private, intrinsic, self-contained: as qualia in the philosopher's sense, as the contents of the inner theater diagnosed in Chapter 2.

On that picture, physical processing seems to face a second job: producing the private exhibit after its world-directed work has been explained. The representationalist challenges the description that gives it this job. Seal an exhibit behind glass and the unreachability you then discover may belong to the glass. But an intrinsic quality need not be nonphysical, and rejecting the exhibit does not refute every theory that posits such a quality. The diagnosis draws on the earlier transparency argument; it cannot settle the identity question by itself.[4]

But strong representationalism (developed across Chapters 4 to 6 from the transparency of experience) contests precisely that description. Phenomenal character is not an inner self-standing item at all. It *is* a way a system represents its environment — the representational content of perceptual and bodily experience, made manifest from the inside. Experience involves the world in its very constitution. It directs itself, by its nature, outward. Physical processes that realize experience are the processes by which an organism represents its environment, and those representations, lived from the inside, simply *are* the experience of that environment. There exists no further inner item to account for, because the supposed inner item never figured in any faithful description of experience to begin with.

Once that description shifts, the gap narrows without vanishing. What looked like a chasm between physical processes and inner phenomenal items depends on a picture already dismantled. Strip the picture away and the question becomes a hard scientific one: how does a physical system represent its environment in the perspectivally rich way conscious organisms do? Levine's quieter epistemic gap survives, as does Chalmers's argument that ideal conceivability carries modal weight. Chapter 11 asks how the identity claim meets both. The glass-case diagnosis removes only the chasm manufactured by treating phenomenal character as intrinsic inner paint.

### 9.3 The Limits of Conceivability: Which Duplicate Has the Lights On

The zombie now meets the standard Chapter 8 applied to Mary. Even after ideal reflection, we can coherently imagine the scenario as a way the world might actually be. Chalmers calls this *ideal primary positive conceivability* and argues that it entails primary possibility.

Apparent possibilities such as water without $H_2O$, heat without molecular motion, and the morning star without the evening star lost their force once inquiry showed what the terms picked out. The conceiving may remain available under some description; what changed was our judgment about what the world could contain. Chalmers has a worked-out reply to exactly these examples: ordinary identities involve contingent ways of fixing reference, he argues, while phenomenal concepts do not. Chapter 11 meets that reply in full.

Grant the conceiving permanently. Conjoin the whole physical story of a duplicate with *and the lights are off*; fill in the scenario coherently, consider it as actual, and no contradiction surfaces under ideal rational reflection. I grant Chalmers that stronger premise, not merely a vivid mental picture. I refuse the next inference: that this coherent conception describes a metaphysically possible world.

The concept you just used has a history of its own. You know what seeing red is like because you have seen red. On the account developed in Chapter 11, an occurring experience anchors a demonstrative, *this kind of seeing*, and a capacity to recognize it again. The concept picks out a world-presenting state, not an inner swatch; what that state presents reaches outward to the tomato. For knowers with our architecture, reading a physical description does not by itself confer that recognitional capacity. Acquiring it and connecting the two concepts are different achievements. Evidence can warrant an identity between what they pick out without making it follow from the concepts alone. This acquisition story helps explain why the gap can persist if the

identity holds. It neither proves ideal conceivability nor shows that every possible knower lacks an a priori route.[5]

And there the argument reaches its real point of dispute. The story about concept acquisition explains why zombie conceivability persists if the identity claim holds. It does not yet tell us what that conceiving reveals about possibility. Chalmers says the phenomenal case differs from water precisely here: for core phenomenal concepts, no contingent description fixes the reference and waits for science to uncover the underlying nature. What verifies the phenomenal description also satisfies it. Ideal conceivability therefore carries modal weight, unless the physicalist accepts a necessity that no a priori conceptual route reveals — the "strong necessity" Chapter 11 meets directly.

Persistent conceivability does not decide between the views, because the type-B physicalist predicts it too. The dualist can still hold that the conceiving tracks a real possibility; the physicalist can deny the governing modal principle. The zombie argument shows that the epistemic gap runs deep and stays. The modal verdict depends on the identity claim of Chapter 6 and the strong-necessity dispute in Chapter 11.

Two duplicates now need to stand side by side. **The full physical duplicate** belongs to a world matching ours in its physical history as well as its present state. It retains the organism's encounters, learning, and content-fixing relations. **The present-role duplicate** matches current internal organization and input-output dispositions but lacks that history. Calling the second a *complete functional duplicate* would hide the omission: a wide functional account includes the very relations we have removed.

On the theory defended here, the first shares our content and phenomenal character; a further stipulation that it lacks experience remains conceivable but cannot hold. The second does not inherit the first's content merely by matching its present roles. Chapter 15 takes up the historyless case as Swampman and accepts the cost of the historical theory. A historyless present-role duplicate may therefore lack felt life: the limited zombie possibility this account allows. One comparison fixes the physical world and its past; the other deliberately does not. The different verdicts follow that difference in the cases.

Neither duplicate describes an existing language model. Actual models have training histories, including causal links to human linguistic practices; they are not historyless. Nor has Part Four claimed a perfect human functional duplicate. Those systems require assessment on their actual histories and capacities, not classification by resemblance to either imagined creature.

### 9.4 The Completeness of the Physical

Even if the glass-case diagnosis and the phenomenal-concept account leave the ideal zombie intact, the argument must still leave the thought experiment and enter this world, where a non-physical mind would have work to do. Here it meets a constraint independent of how zombies seem.

Princess Elisabeth of Bohemia spotted the trouble at its source. In 1643, at twenty-four and exiled in The Hague, she asked Descartes how a soul with no extension, surface, or location could push

the body's animal spirits.[6] Pushing seemed to require contact, and contact extension—the very feature his soul lacked. His appeal to notions proper to soul, body, and their union did not explain the interaction. Elisabeth replied that she would sooner grant the soul extension than grant an immaterial thing power to move a body. Descartes finally suggested that people who do not philosophize understand the union best—advice that concedes more than it answers. Nearly four centuries of physics have added a harder question: whether anything remains for the soul to do.

The modern question concerns *causal completeness*: do prior physical conditions suffice for physical effects, fixing either their outcomes or their chances? The case comes from the success of physiological explanation. Retinal activity, nerve transmission, and muscle contraction yield to accounts in terms of physical mechanisms. No established result in that record requires a non-physical intervention. Conservation laws constrain proposed mechanisms, but cannot establish completeness on their own: an intervention need not advertise itself as newly created energy.[7]

David Papineau traces the historical case from early mechanics, through the unsettled status of forces, to conservation and twentieth-century physiology. It supports a general conclusion beyond the individual mechanisms examined. That makes completeness a serious empirical commitment, open to contrary evidence, rather than a theorem that forbids finding any.

For a dualist who treats consciousness as distinct from its physical basis, three options remain.

1. **Change the physical course.** A mental cause could alter a synapse or change the chances of an outcome beyond what prior physical conditions fix. That rejects completeness and carries an empirical burden: identify the alteration and show why the physical account cannot explain it.
2. **Cause the same effect twice over.** A mental cause and a distinct physical cause might each suffice for the same movement. Philosophers call this *overdetermination*. The mental cause would genuinely act, even though the physical cause already sufficed; it would not become causally idle. Completeness alone cannot rule this out. Rejecting it requires the further premise that ordinary mental effects do not systematically have two distinct sufficient causes. The dualist can make the causes depend on each other by law, avoiding an accidental coincidence. My objection is comparative: without independent evidence for that double causation, identity preserves mental efficacy and the physical explanation without adding laws that keep distinct causes aligned.
3. **Accompany without influencing.** Consciousness could occur alongside the physical process while causing no physical effect. Thomas Huxley proposed this in 1874 through his steam-whistle analogy: the whistle accompanies the locomotive's work without driving it.[8] The position, *epiphenomenalism*, also appears in Jackson's original Mary paper. It preserves completeness, but conscious experience then contributes nothing to the causal production of our reports about it.

The third option exposes an evidential problem, though it does not refute property dualism. If the phenomenal property never helps cause a judgment about experience, that judgment cannot count as a causal response to it. The dualist's common-cause reply holds that the neural state produces the judgment and fixes the phenomenal property in one stroke. Across worlds governed by the same psychophysical laws, that arrangement may secure ordinary reliability. The question shifts

to warrant: how does the judgment become about a property omitted from its causal production? Chalmers answers by making acquaintance constitutive rather than causal, and Chapter 11 gives that answer its full hearing. This problem belongs to epiphenomenalism; it cannot simply be transferred to the overdetermination option, which retains mental causes.

Causal completeness concerns the sufficiency of physical causes, not everything that exists or which vocabulary explains best. A property dualist can grant it. Nor does it make particle physics the best language for understanding a mind: bodies representing their environments remain wholly physical without demanding translation into physics-talk. The constraint operates inside nature and stays silent about whether nature has an author.

The case against a distinct mental cause therefore needs more than completeness. Interactionism must challenge the physiological evidence; systematic overdetermination must justify its paired causes; epiphenomenalism must explain our warrant for judgments about a causally silent property. I favor identity over those alternatives, but the reasons differ. Panpsychism and Russellian monism need not posit a second cause at all: they locate phenomenal nature in what already does the physical work. Chapter 10 meets that rival on its own ground. The claim that mind consists in software likewise proposes no cause outside the physical order; Chapter 20 gives it a separate hearing.

### 9.5 What Dualism Gets Right

Dualism persists not because philosophers are confused but because it captures something real.

1. It gets right that experience has a first-person perspective. Experience happens *for* someone, *to* someone. The bat's echolocation presents the world to the bat. The bat occupies a point of view. This first-person character cannot be reduced away without loss, and any adequate account of consciousness must preserve it.
2. It gets right that objective description leaves something out — not something mysterious over and above the physical, but something perspectival. The third-person vocabulary of physics and neuroscience describes the world from no particular point of view. Experience has a point of view. These are different modes of description, and switching between them involves a genuine shift, not just a change of vocabulary. Nagel was right about this.
3. It gets right that consciousness matters. That the world contains subjective experience (that there is something it is like to be a bat, or a human being, or presumably many other organisms) is not a trivial addendum to the physical story. A physicalism that explains it away has changed the subject, not won the argument.

These observations can tempt us into shortcuts: experience has a point of view, so it must belong to a separate realm; objective description leaves the perspectival out, so the perspectival must be nonphysical; consciousness matters, so physics cannot account for it. None supplies the missing inference. Chalmers's developed argument takes a harder route. He offers a modal principle connecting ideal conceivability to possibility, and the dispute concerns that principle's warrant, not whether he forgot to provide one.

On the account defended here, first-person character belongs to a physical organism presenting its environment from its particular embodied position in the world. A detached description

can describe that standpoint without giving its reader the standpoint itself. That limit does not establish a second realm. Consciousness matters because experience engages the world. Whether this physicalist account succeeds depends on the identity and modal arguments already named, not on dismissing the observations dualism preserves.

### 9.6 Living With the Residue

Even after all this, a genuine puzzle remains: why should any arrangement of matter represent the world *to itself*? Why should there be any inside view? I don't think anyone has answered it fully. But once the inner-theater picture goes, the question concerns why a perspectival relation to the world should feel like anything at all, not how physical processes produce inner phenomenal items. Chapter 11 asks whether that residue is the predicted shadow of an identity rather than the unpaid debt of an incomplete theory.

One limit would remain even after a scientific answer: no third-person account can hand its reader the first-person standpoint it describes. A map can locate a viewpoint without putting you there. That marks a limit on transmission and access, not necessarily an unpaid explanatory fact.

Dualism points at something real, and physicalism, handled carefully, can honor what it points at without following it into a second ontological realm.

Something it is like to be you, reading this: real, physical, directed at the page, and thoroughly of this world. The bat is welcome to its own version.

### Chapter Summary

**The claim.** Nagel's bat, Chalmers's hard problem, and the zombie give the dualist intuition real force. The lasting first-person gap does not by itself prove dualism; Chalmers's argument needs the further modal principle this book contests. Unequal routes of access remain compatible with one physical fact if the identity holds.

**Where the argument stands.** Ideal conceivability remains compatible with the type-B account; its modal force remains disputed. Causal completeness burdens interactionism but permits systematic double causation or causally silent experience. Identity avoids the former's extra coordination, while the latter owes an account of our knowledge of a property that never helps cause our reports.

**What's left open.** The modal dispute remains real. Does ideal conceivability reveal possibility, or can an a posteriori identity outrun it? The identity claim must carry that answer; causal closure cannot.

**The hand-off.** A subtler rival accepts closure and locates phenomenal nature in matter's own causal powers. No second cause need enter. The next chapter asks what that proposal explains.

### Footnotes

**1.** Thomas Nagel, "What Is It Like to Be a Bat?", *Philosophical Review* 83 (1974): 435–450. Nagel's paper presents not an argument for dualism but a case for the irreducibility of subjective facts to objective description. His conclusion remains agnostic about the metaphysics: he does not claim that consciousness lies outside the physical, only that our current conceptual resources cannot

explain the relationship. At 446 he expressly declines to conclude that physicalism must be false; this does not remove his challenge to existing reductive accounts. In *The View from Nowhere* (New York: Oxford University Press, 1986), Nagel criticizes physicalist and functionalist objectification and calls for a transformed conception of reality that gives subjectivity a fundamental place. The reconciliation with physicalism proposed here belongs to this book, not to Nagel's later position. The point about the first-person perspective as a real feature of consciousness that objective description cannot fully capture remains correct and important — the question, taken up below, concerns whether it supports dualism or merely reveals a limitation intrinsic to third-person description. Strong representationalism offers a way to accommodate Nagel's concern: experience *is* constitutively perspectival (it presents the world *to* an organism), and that perspectival character obtains within nature without requiring any non-physical ingredient.

**2.** Joseph Levine, "Materialism and Qualia: The Explanatory Gap," *Pacific Philosophical Quarterly* 64 (1983): 354–361. Levine introduces the explanatory gap as a gap in our understanding rather than in the world, leaving open the possibility that physicalism remains true even though we cannot see *why* a given physical process gives rise to a given phenomenal character. Chalmers's Hard Problem asks why and how physical processing is accompanied by experience; his arguments for ontological consequences add a further step (footnote 3). The explanatory question and the proposed metaphysical answer must remain distinct.

**3.** David Chalmers, "Facing Up to the Problem of Consciousness," *Journal of Consciousness Studies* 2 (1995): 200–219; and *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press, 1996), esp. chs. 3–4. Chalmers gives Levine's gap ontological force through zombie and conceivability arguments: if the complete physical truth can hold while phenomenal truth varies, physicalism is false. The bridge from an epistemic situation to an ontological conclusion is therefore a modal argument, not a relabeling. This chapter contests that bridge while preserving the gap it begins from; the diagnostic move appears in footnote 4.

**4.** The diagnosis targets an inner-theater rendering of the Hard Problem, not every account of intrinsic phenomenal properties. Its representationalist rejection of a private exhibit does not show that intrinsicality entails nonphysicality, nor does it dispose of Levine's explanatory challenge or Chalmers's modal argument. Michael Tye, *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind* (Cambridge, MA: MIT Press, 1995), esp. ch. 5, and Fred Dretske, *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995), esp. ch. 3, develop the identity version of intentionalism on which phenomenal character *is* (a specific kind of) representational content — not merely supervenes on it or is necessitated by it, but is numerically identical to it. This book endorses that identity claim. On the strong intentionalist view, the apparent gap between physical processes and phenomenal properties narrows to a more tractable gap between physical processes and *world-directed representational content* — the latter itself a physical, naturalizable phenomenon. The identity claim supplies the book's answer to the zombie argument if its necessity holds: if phenomenal character just *is* representational content of the right embodied, world-directed kind, there exists no further phenomenal ingredient a duplicate could lack once its content is fixed. What fixes it is not the duplicate's present physical constitution alone. A duplicate sharing our physical processes *and* the world-involving history that gave those processes their

content — which is what Chalmers's zombie is, since his world duplicates ours in its whole physical past — shares our representational contents, and therefore our phenomenal character. Chapter 11 defends the necessity that this answer requires. Strip the history and the entailment lapses; Section 15.6 takes up that case in Davidson's Swampman and concedes it openly.

**5.** The conceivability-to-possibility inference receives critical examination in Brian Loar, "Phenomenal States," *Philosophical Perspectives* 4 (1990): 81–108, where Loar argues that phenomenal concepts form a special class of recognitional concepts that directly pick out physical states, explaining the conceivability of zombies as an epistemic phenomenon without ontological significance. See also Christopher Hill, *Sensations: A Defense of Type Materialism* (Cambridge: Cambridge University Press, 1991). Chalmers's restricted principle is that *ideal primary positive conceivability entails primary possibility*: "Does Conceivability Entail Possibility?", in *Conceivability and Possibility*, ed. Tamar Szabó Gendler and John Hawthorne (Oxford: Oxford University Press, 2002), 145–200, esp. secs. 1–3 and 5–6. Positive conceivability requires coherent modal imagination of a scenario that verifies the statement; ideality requires that the imagining survive ideal rational reflection; primariness evaluates the scenario as actual. The acquisition account helps explain conceptual asymmetry given the identity; it does not independently establish ideal conceivability, the identity, or the failure of Chalmers's modal principle. Nor does it establish that every possible knower lacks an a priori route. Chapter 11 develops the strong-necessity standoff, including Chalmers's dilemma against phenomenal-concept strategies; Section 9.4 then adds a constraint that stands even for a reader who grants the conceiving and refuses the diagnosis. The identity case belongs to Chapter 6, and this chapter neither strengthens nor substitutes for it.

**6.** The exchange runs through 1643: Elisabeth's opening letter of 6 May poses the question how an immaterial, unextended soul can determine the motion of the body's animal spirits; Descartes' reply of 21 May distinguishes the primitive notions of soul, body, and union; Elisabeth's letter of 10 June answers that she would find it easier to attribute matter and extension to the soul than to grant an immaterial being the capacity to move a body and be moved by it; and Descartes' reply of 28 June directs her to everyday life — those who never philosophize, he suggests, grasp the union of soul and body best. *The Correspondence between Princess Elisabeth of Bohemia and René Descartes*, ed. and trans. Lisa Shapiro (Chicago: University of Chicago Press, 2007). Long filed as a footnote to Descartes, Elisabeth now receives treatment as a philosopher in her own right; Shapiro's introduction makes the case.

**7.** David Papineau, *Thinking About Consciousness* (Oxford: Clarendon Press, 2002), Appendix: "The History of the Completeness of Physics," supplies the historical case for completeness as an empirical commitment rather than a metaphysical stipulation. His causal argument separately requires mental efficacy, physical completeness, and the rejection of systematic overdetermination; see ch. 1, secs. 1.2 and 1.5, esp. 26–28. Section 1.5 considers the proposal that distinct sufficient causes remain counterfactually dependent and rejects it on defeasible grounds of theory choice, not by identifying overdetermination with epiphenomenalism. In the probabilistic formulation, prior physical conditions fix the chances of physical effects; an interactionist who proposes different chances thereby disputes completeness. Establishing such a dispute empirically would require a detectable, reproducible departure from a supported physical account, not merely an unexplained

result. Friedrich Beck and John Eccles, "Quantum Aspects of Brain Activity and the Role of Consciousness," *Proceedings of the National Academy of Sciences* 89 (1992): 11357–11361, offer an interactionist proposal. Jaegwon Kim, *Mind in a Physical World* (Cambridge, MA: MIT Press, 1998), esp. ch. 2, develops the exclusion problem: given sufficient physical causes, distinct mental causes require an account of their contribution. Identity avoids competition between distinct causes. For resistance to the causal argument's reductive deployment, see Tim Crane, "Mental Substances," in *Minds and Persons*, ed. Anthony O'Hear (Cambridge: Cambridge University Press, 2003), 229–250. Conservation laws alone do not establish closure or rule out every form of interaction; Papineau's historical argument appeals to physiological mechanisms as well. The relevant empirical scale here is nerve and muscle, not global energy accounting in general relativity.

**8.** Thomas Henry Huxley, "On the Hypothesis that Animals Are Automata, and Its History," *Fortnightly Review* 16 (1874): 555–580. Huxley's steam-whistle analogy makes consciousness a collateral product of the body's working, without power to modify it. Frank Jackson, "Epiphenomenal Qualia" (cited in Chapter 8), stages the complaint in an objector's voice — qualia "do nothing, they explain nothing, they serve merely to soothe the intuitions of dualists" — and then embraces it: "This is perfectly true; but is no objection to qualia." Jackson's later recantation is told in Chapter 8. The property dualist's reply to the resulting epistemic worry appeals to a common physical cause — the neural state that produces a phenomenal judgment also fixes the phenomenal property, so the two correlate reliably even though the property does no causal work. David Chalmers formulates the difficulty as the "paradox of phenomenal judgment" and answers it by grounding knowledge of consciousness in acquaintance rather than causal responsiveness: *The Conscious Mind* (New York: Oxford University Press, 1996), ch. 5. Chapter 11 takes up the paradox and the acquaintance reply in full.

## Chapter 10: The Case Against Panpsychism

> *"One thing we know about physical stuff, given that (real) physicalism is true, is that when you put it together in the way in which it is put together in brains like ours, it regularly constitutes—is, literally is—experience like ours."*
>
> — Galen Strawson, "Realistic Monism: Why Physicalism Entails Panpsychism"

---

CHAPTER OVERVIEW

Could matter's own nature explain both what it does and how we feel? Some forms of panpsychism pursue that aim without adding a second cause to physical events or asking many tiny subjects to merge into one of us. I take that gain seriously, then ask why the qualities we feel must come from qualities at the base of nature. Knowing pain from within tells us how it feels, but need not tell us what the most basic parts of matter contain. My own view says that felt character consists in how we present the world, a claim we must learn from evidence rather than prove from concepts alone. I favor it because the perceptual cases give it explanatory work to do without a further felt basis; pain's badness and mood still need accounts of their own.

### 10.1 The Disciplined Alternative

Panpsychism offers an alternative that the case against a second mental substance does not reach. It locates consciousness within matter itself.

Bertrand Russell noticed that physics tells us how matter *behaves*—its structure, relations, and effects—without disclosing its complete concrete nature.[1] It characterizes an electron through mass, charge, and other relations or dispositions. Follow physical description down and it ends in more structure and behavior. A blank may remain where defenders of the view place matter's *intrinsic natures*, or *quiddities*.

Now add a second thought. Exactly one piece of the world spares us that ignorance, one place where we live on the inside of the stuff rather than reading its behavior off from outside: our own experience. I do not infer my headache from its effects on my behavior. I have it. Now for the elegant move.

> Physics leaves the intrinsic natures of matter blank; consciousness is the one intrinsic nature we know directly; so perhaps consciousness, or some humbler proto-conscious ingredient, fills the blank. Perhaps the intrinsic nature of the physical, the categorical stuff underlying all that structure, is experiential. Matter, all the way down, has an inside.

*Russellian monism* names the family, not a cautious cousin of panpsychism. Russellian panpsychism says the categorical properties realizing physics's causal roles are phenomenal. Russellian panprotopsychism puts non-experiential but consciousness-constituting properties in that role.[2] Philip Goff develops this role-and-realizer route: physics tells us what mass or charge does, while experience supplies our one positive clue to what the realizer might be.

The strongest version need posit neither a spectator in every particle nor a hidden quality that does no causal work. A *powerful quality* has its qualitative nature and causal power inseparably: one property, described in two ways. A *panqualityist* begins with qualities without microscopic subjects and asks how their organization constitutes a conscious whole. Section 10.7 returns to these proposals. Their attraction comes from unification: the same nature could explain both what matter does and how experience feels. The question is whether they can earn that explanation, not whether small subjects sound peculiar.

Galen Strawson, a British philosopher known for refusing any theory that makes experience less real than physics, arrives by a different road. His "real physicalism" begins from two claims:

1. experience is unquestionably real; and
2. experience is wholly physical.

He then denies that experiential being could radically emerge from wholly non-experiential being. So experience, or something of its nature, must already belong to the physical foundations.[3] Goff selects a phenomenal candidate for physics's hidden realizers; Strawson argues that real physicalism itself cannot begin with the wholly unfelt. The arguments meet at fundamentality, but the book resists different steps: Goff's selection of the phenomenal realizer and Strawson's move from no derivation to no emergence.

The appeal, once you feel it, sticks. Panpsychism promises to dissolve the hard problem rather than solve it. The hard problem asks how mere matter, which has no inside, could ever produce an inside—how the water of the physical becomes the wine of the felt. Panpsychism answers that it never does, because the inside was there all along. No miraculous switch-on occurs in evolutionary or developmental history; the lights were never off. To a reader who rejects dualism but cannot believe consciousness just appears when matter gets complicated enough, this lands as a relief. David Chalmers has mapped the options, Philip Goff has defended the view for both general and specialist readers, and it has claimed a convert this book cannot wave away.[4]

And the stakes reach past the seminar. In some AI talk, panpsychist language merges with the thought that sufficiently integrated information processing has an intrinsic felt side. That conjunction would change the machine question, but panpsychism alone does not settle when micro-level phenomenality forms a machine-level subject, or whether it ever does. The view can put consciousness at the bottom and still owe a threshold at the level of the machine. Part Four's question survives.

### 10.2 Why the Hammer From the Last Chapter Misses

Chapter 9 gives interactionist dualism an empirical burden and systematic double causation a separate explanatory cost. A Russellian who identifies phenomenal nature with the physical cause need accept neither. Causal completeness does not decide this dispute.

The Russellian versions at issue accept causal closure. They posit no second substance or ghostly intervener. Their phenomenal properties add no forces; they supply the *intrinsic natures of physical properties themselves*, the categorical ground beneath the dispositions physics describes. Nothing new pushes anything. The electron behaves exactly as physics says while having, or deriving from, some inner phenomenal nature. The completeness hammer passes through because these versions never claimed a causal gap.

These versions dispute no equation, closure principle, or conservation law. They ask what the equations are *equations of*. The argument therefore turns on what explains experience and what its felt sharpness shows. Michael Tye supplies the hardest internal test: his earlier representationalism shares the identity claim, while his later panpsychism adds the fundamental phenomenal feature this book rejects.

### 10.3 The Convert in the Witness Box

Michael Tye spent thirty years developing a view this book engages deeply: phenomenal character consists in representational content of the right kind. His conversion matters because he keeps much of that account while arguing that it needs a fundamental phenomenal basis. He announced the change in a 2021 book and explained it again in a 2024 essay for a general audience.[5]

The late view develops across two formulations. In Tye's 2021 book, a sharp fundamental determinable, *consciousness**, grounds phenomenality while ordinary determinate consciousness remains tied to representational content and may inherit functional vagueness. His 2024 restatement presses the essential sharpness of phenomenal consciousness more directly. Try to imagine a state for which there is genuinely no fact whether anything is felt, he says, and you cannot. The lights are on, however dimly, or off. We know this, he says, a priori.[6]

Every familiar *physical* candidate for consciousness is vague: neurons, oscillations of "approximately 40 hertz," functional and behavioral states, all shot through with borderline cases. If the phenomenal base is sharp, it cannot be one and the same as any such vague candidate, and a dilemma follows.

1. Sharp consciousness *emerges*, brutely, the moment the vague physical crosses some threshold (something "radically different and new and non-physical and sharp" popping into being, exactly the miraculous switch-on the whole tradition has found incredible); or
2. it was there all along, fundamental, at the base.

Tye takes the second horn and becomes a panpsychist not despite his representationalism but, he thinks, in continuity with it.

His answer to the combination problem begins by denying that rich micro-experiences must assemble into our experience. At the bottom sits the bare determinable *consciousness**; representational content supplies the determinate character. A transfer condition then specifies when the

feature belongs to an organized whole. Section 10.6 tests that proposal rather than asking him for one he has already supplied.

Tye has not abandoned representationalism; he argues that, followed to its end, it requires panpsychism. If representationalism and the sharpness premise hold, fundamental consciousness seems to follow.

### 10.4 The Dispute Over Sharpness

The book's objection treats the sharpness argument as the conceivability problem in another form. From a fact about *us* — that we cannot imagine a borderline phenomenal case — the argument reaches a fact about *the world*: that the phenomenal base is sharp in its own nature. A recognitional concept deployed from the inside like a demonstrative may present its referent as all-or-nothing whether or not the worldly kind it tracks is vague, the way a thermometer that reads only "hot" and "cold" tells you nothing about how many temperatures exist. The unimaginability is real. The disputed step asks what it reveals: the grain of consciousness itself or the instrument with which we approach it. If the second diagnosis is right, Tye's dilemma between brute emergence and bedrock consciousness loses its first premise without any help from panpsychism.

Tye has faced the concept–property objection in print. He credits Michael Antony with a view on which consciousness could be vague despite our sharp present concept.[7] Tye answers that sharpness belongs to consciousness itself; he takes that verdict to disclose part of its essence a priori. The answer does not supply a neutral bridge from concept to property, but it locates the disagreement honestly. One side treats the first-person verdict as insight into the property's nature; the other treats the same all-or-nothing appearance as the expected profile of an acquaintance-grounded recognitional concept. That appearance alone does not settle which explanation is right.

The standoff does not decide the chapter. Grant Tye the fundamental feature and test the mechanism he gives it: transfer from microstates to a suitably organized whole.

### 10.5 The Problem That Has Dogged the View Since James

Constitutive Russellian panpsychism takes rich phenomenal properties as fundamental and holds that they, with physical structure, ground macro-consciousness. Its panprotopsychist sibling avoids micro-subjects but still owes the passage from non-experiential proto-properties to experience. William James identified the panpsychist problem in 1890, before the modern view had a name.

James put it with characteristic bluntness: take a hundred feelings, shuffle them and pack them as close together as you can; they will never, by any squeezing, become one feeling. "Each remains, in the sum, what it always was." A row of people can each feel a fraction of a sentence; you do not thereby get a person who feels the whole sentence. Feelings do not sum.[8] If particles carry dim experiences, my unified visual field must somehow be built from experiences belonging to many micro-subjects. Yet my experience and yours do not, by adjacency, fuse into a third experience belonging to someone who feels both.

Chalmers separates the debts:

1. how micro-*subjects* yield a macro-subject;
2. how micro-*qualities* yield familiar qualities; and
3. how microexperience yields the structure of our experience.[9]

Goff's phenomenal-bonding proposal answers the first at full strength. He grants that micro-subjects, merely existing with their experiences, cannot necessitate a further subject. Add a relation, however, and the entailment may change: if subjects stand in a special phenomenal-bonding relation, that relational fact could necessitate a macro-subject whose experiences are jointly co-conscious. We should not expect transparent access to this relation, Goff argues, because introspection presents only one subject and physics presents only mathematical-causal structure. He even allows that an empirically known relation, under another conception, might be the bond; his concrete candidate is the spatial relation.

Obscurity alone cannot settle the matter; the book also accepts an a posteriori necessity it cannot derive. The sharper question asks whether the book's own integration criterion already hands Goff the relation he needs. It does not. Integration here locates a candidate bearer — the control organization to which capacities belong — while phenomenal bonding adds the further claim that the same relation necessitates a numerically new subject and one field of co-conscious experience. That extra necessitation needs support.

Bonding therefore meets James, but it does not yet discharge Chalmers's full bill. A spatial bond yields unrestricted composition—your nose, Venus, and my left shoe acquire a group subject—or else needs a principled restriction. A restricted bond still owes quality and structure. The crack has not merely moved; Goff has proposed mortar. We still need to know why it holds only these bricks and builds this room.

Cosmopsychism reverses the direction, beginning with one cosmic subject and carving local minds out of it. Goff's hybrid version avoids upward subject-summing and pays for it elsewhere: a cosmic subject, strongly emergent local subjects, transfer, and laws of localization and thinning fitted to the distribution of consciousness we actually observe.

Strawson's no-radical-emergence argument remains distinct. He and this book agree that experience literally consists in physical goings-on. He infers that experiential being must occur in the ultimate base because the wholly non-experiential cannot yield it. The book denies that inference: if phenomenal character consists in world-directed representational activity, nothing phenomenal emerges *from* that activity. One fact falls under two concepts that do not entail each other. Chapter 11 owns the price—a strong necessity without an a priori derivation. Strawson's challenge survives as pressure on that identity, not as a role/realizer blank Goff gets to fill.

The verdict remains comparative. If the identity claim holds, phenomenal fundamentals lose their motive while bonding, transfer, and decombination retain theirs. If it fails, Strawson's pressure returns. Combination alone proves neither result.

### 10.6 The Idle Wheel

Tye designed an ingenious move to make all of this go away. At the most fundamental level, he says, there are no tiny experiences of red or pain to combine. There is only consciousness*, a sharp

determinable with no species. At the macro level, a conscious state consists of that determinable plus representational content, which supplies the determinate character — red rather than fear, thirst rather than silence. Consciousness* alone does not make a state conscious.

How, then, does a brain get it while a rock does not? Tye's official answer is *transfer*. Consciousness* passes from microstates to a complex whole once that whole is definitely organized as a relevant global-workspace state, its content widely available to reporting, reasoning, memory, and planning. He compares the transfer to graceful dancers forming a graceful ensemble, or meaningful words forming a meaningful sentence. This blocks the simple switch-on objection: nothing phenomenal gets created from nothing at the macro level. The fundamental determinable acquires a new bearer, and content gives the resulting state its specific feel.[10]

The analogies make a legitimate point: a property of parts can belong to an appropriately organized whole. No literal migration need occur. But the gracefulness of an ensemble depends on how the dancers move together, and the meaning of a sentence on how its words combine. We still need the corresponding explanation for consciousness*. Broadcasting admits borderline cases, so Tye requires the workspace condition to be *definitely* met. That specifies his proposed condition; it does not yet explain why satisfying it makes the whole bear a fundamental phenomenal feature. Transfer blocks the charge of creating phenomenality from nothing while leaving a constitutive connection to justify.

Consciousness* therefore has a stated job in Tye's theory: supplying *that* there is phenomenality while content supplies *which* phenomenality. If sharpness independently requires it, the feature earns theoretical work. The challenge concerns whether that argument and the transfer proposal explain more than an account without a fundamental phenomenal determinable.

The rival can answer in terms our own view must respect: *You accept an identity whose necessity you cannot derive. Why may my constitutive connection not end explanation too?* It may. An unexplained necessity does not disqualify one theory while passing unnoticed in another.

Fair comparison needs one stopping rule for every theory. An identity or basic law may end explanation when:

1. evidence or abductive support extends beyond the bare fact that the target phenomenon exists;
2. it unifies a wider pattern or risks a discriminating failure rather than renaming the missing step; and
3. a rival does not recover the same pattern with fewer fundamental properties or laws.

Failing this test does not prove a view false. It denies that view explanatory credit.

Apply the rule to the same case. When a page looks blurred, the account in Chapter 6 connects the changed appearance with a less determinate presentation of its edges. Stimulus conditions, discrimination, and correction constrain that description apart from the report of how it looks. The gain lies in using the same profile to explain what appears, how it can mislead, and what changes when the presentation sharpens. The identity claim says this profile constitutes the character. A theory of fundamental qualities can preserve all those explanations, but it still needs to show what its qualitative basis contributes beyond them. Neither the blur comparison nor its

clinical counterparts prove identity over necessary grounding. The narrower advantage is that the representational account already organizes the evidence before the further basis enters. That advantage reaches the perceptual cases developed here. Pain's badness and mood remain unfinished; an extra fundamental quality gains no support merely from that failure, but our account gains no explanatory credit for them either.

Phenomenal bonding passes only if the bond earns standing beyond the combination result it was introduced to secure. Identifying it with space predicts the unrestricted group subjects Goff openly accepts; that keeps the proposal consistent at a steep counterintuitive cost. Identifying it with an integration relation yields a better boundary, and the panpsychist can treat organization and phenomenal bonding as two conceptions of one relation. The remaining burden is comparative: what independently supports that categorical identification and its necessitation of a new subject, and what pattern would it explain or predict beyond the public organization? Bonding remains possible. On present evidence it receives no explanatory bonus.

Tye's transfer fares similarly. Global availability has independent empirical credentials as a candidate condition for state consciousness. A law making consciousness* change bearers exactly when that condition becomes definite does not. It marks the sharp crossing the theory needed. Goff's localization and thinning principles borrow candidate boundaries and information structures from neuroscience, then add laws that create local subjects and select which cosmic phenomenal properties they inherit. Those laws can be made precise and may someday risk a distinct prediction. As sketched, they reproduce the known distribution while adding a cosmic subject, property transfer, and two selections.

Identity thus accepts a strong a posteriori necessity; bonding accepts a constitutive relation; Tye accepts consciousness* and its transfer; hybrid cosmopsychism accepts a cosmic subject and principles for local subjects. Counting names cannot choose between them. The question is whether the added commitments explain a pattern the independently constrained representational account misses. At present I find no such gain. A well-supported sharp boundary, a better account of unity, or a phenomenal contrast with our complete profile fixed could change the comparison.

### 10.7 The Strongest Russellian Redoubt

The powerful-quality proposal introduced in Section 10.1 deserves that same comparison. Think of pain. The episode hurts, captures attention, and moves the organism to protect itself. The rival proposes one concrete property that does both jobs. Its qualitative nature need not float free of its powers, and a panqualityist can develop the proposal without microscopic subjects. The theory seeks a unified basis for causation and character, not an extra cause alongside a complete physical one.

The further step goes from that organism-level episode to qualities in the fundamental base. Acquaintance alone does not supply it.[11] It reveals a manifest profile — how red appears, how pain hurts — without settling whether the profile consists in relations or inherits its character from more basic qualities. The Russellian can offer an inference rather than claim direct revelation: qualitative continuity might make the whole more intelligible. But that inference needs a reason why relational organization cannot constitute the character, or a positive explanatory gain from

the proposed continuity. Merely requiring a concrete basis for a power does not establish that the basis must be phenomenal or specially fitted to constitute phenomenality.

A local Russellian stops at the organism, and that version sidesteps the descent entirely. If *categorical* means only the concrete realization of a relational profile, the local view converges with the identity claim defended here. A local Russellian may instead posit a necessarily co-instantiated categorical realizer — a view that stays metaphysically distinct even if no phenomenal contrast ever appears, and that earns no abductive advantage without independent support or some further explanatory payoff. Chapter 6 sets the guard that keeps that test honest: content, mode, determinacy, field structure, and attention receive specifications independent of the phenomenal verdict, so no after-the-fact enlargement of the profile counts as a rescue.

The blank in physics gives the Russellian room. It does not give the view title, and it does not make the room's occupant fundamental.

My verdict remains categorical: nothing mental or phenomenal is fundamental. Its warrant remains comparative. Part Two's world-presenting account explains the contrasts while preserving their felt reality; the strongest rival can match that account but has not yet shown why a fundamental qualitative basis improves it. If the identity fails, or the rival earns a further explanatory gain, this judgment must answer to that result. For now, consciousness gives us reason to investigate what organized matter can do before deciding what every fundamental part must contain.

**Chapter Summary**

**The claim.** Panpsychism says experience lies at the base of nature. Panprotopsychism instead puts a nonconscious base there with a special role in forming experience. Both read the explanatory gap as evidence for such a basis, but current evidence does not make either view the best account.

**Where the argument stands.** Causal closure does not rule out Russellian monism; powerful-quality views give their basis a causal role, and phenomenal bonding offers a way to join subjects without yet explaining the bond's quality, structure, or limits. Knowing experience from within does not reveal a felt basis in basic matter. I favor identity on the perceptual evidence, not on the unfinished accounts of pain and mood.

**What's left open.** Experience may still depend on a concrete property that realizes those relations but remains distinct from them. A stable phenomenal difference with the full content-under-mode profile held fixed would defeat identity. New explanatory gains could also change the comparison.

**The hand-off.** Perhaps the gap never marked missing furniture. The next chapter asks who drew it rather than what could fill it.

**Footnotes**

**1.** Bertrand Russell, *The Analysis of Matter* (London: Kegan Paul, 1927), esp. chs. 14, 24, and 37–38. Russell's point is that physics describes the world purely in terms of structural and relational properties — what entities do, how they are arranged, how they affect one another — and is silent on the intrinsic (categorical) properties that bear those relations. The contemporary revival of this observation, and the term "Russellian monism" for views that fill the intrinsic-nature blank

with the phenomenal or proto-phenomenal, traces principally to David Chalmers, "Consciousness and Its Place in Nature," in *Philosophy of Mind: Classical and Contemporary Readings*, ed. David Chalmers (Oxford: Oxford University Press, 2002), where the position is catalogued as "type-F monism." See also Daniel Stoljar, "Two Conceptions of the Physical," *Philosophy and Phenomenological Research* 62, no. 2 (2001): 253–281, for the structural/categorical distinction that does the work here.

**2.** The distinction between the dorm-room caricature and the disciplined contemporary view matters for fairness. *Russellian monism* is the family: Russellian panpsychism makes the categorical bases of fundamental physical roles phenomenal, while Russellian panprotopsychism makes them nonphenomenal but specially apt to constitute consciousness. Neither attributes *thoughts*, *beliefs*, or rich human-style consciousness to particles. The canonical taxonomy is David Chalmers, "Panpsychism and Panprotopsychism," in *Consciousness in the Physical World: Perspectives on Russellian Monism*, ed. Torin Alter and Yujin Nagasawa (Oxford: Oxford University Press, 2015), 246–276. Philip Goff states the role/realizer formulation and explicitly treats the panpsychist and panprotopsychist variants as the two branches of Russellian monism in "How Exactly Does Panpsychism Help Explain Consciousness?" *Journal of Consciousness Studies* 31, nos. 3–4 (2024): 56–82. Since proto-phenomenal properties are not themselves micro-subjects, panprotopsychism sidesteps subject-summing; but it remains combinatorial by definition and still owes an account of how non-experiential properties constitute phenomenal ones. It also faces the epistemic-selection objection and, conditional on the identity theory, the charge that its base duplicates work already done by organization and content.

**3.** Galen Strawson, "Realistic Monism: Why Physicalism Entails Panpsychism," *Journal of Consciousness Studies* 13, no. 10–11 (2006): 3–31; reprinted with Strawson's replies to critics in *Consciousness and Its Place in Nature: Does Physicalism Entail Panpsychism?*, ed. Anthony Freeman (Exeter: Imprint Academic, 2006). Strawson's core contention is that genuine ("realistic") physicalism must take experiential phenomena to be wholly physical, and that since the experiential cannot be intelligibly derived from the wholly non-experiential — "brute emergence" he regards as unintelligible, not merely unexplained — the experiential must be present among the fundamental physical features of the world. The argument is only as strong as the step from that underivability to fundamentality. Section 10.5 denies the inference rather than the premise: on the identity claim nothing emerges, so physicalism owes no emergence story for the unintelligibility verdict to condemn — phenomenal character *consists in* world-directed representational activity rather than arising from it. For Goff's direct challenge to such a posteriori identities through the transparency and partial revelation of phenomenal concepts, see Philip Goff, "A Posteriori Physicalists Get Our Phenomenal Concepts Wrong," *Australasian Journal of Philosophy* 89, no. 2 (2011): 191–209. That objection concerns the epistemic residue developed in Chapter 11; it does not decide the comparative cost here. The narrower idle-wheel objection is reserved for Tye, whose representationalism supplies the premises it requires.

**4.** For the careful contemporary defense aimed at a general readership, see Philip Goff, *Galileo's Error: Foundations for a New Science of Consciousness* (New York: Pantheon, 2019); for the technical development, Goff, *Consciousness and Fundamental Reality* (New York: Oxford University Press,

2017). Chalmers's role is more cartographic than partisan: he treats Russellian monism as one of a small number of live options and has done more than anyone to expose its central difficulty (see n. 9), which is why his mapping is trustworthy — the most rigorous taxonomy of the view comes from someone drawn to it yet unwilling to shield it from its hardest problem.

**5.** Michael Tye, *Vagueness and the Evolution of Consciousness: Through the Looking Glass* (Oxford: Oxford University Press, 2021), where the conversion is announced and the sharpness argument developed at book length; and "How I Learned to Stop Worrying and Love Panpsychism," *Journal of Consciousness Studies* 31, no. 9–10 (2024): 10–28, the restatement for a general audience that the main text follows. The biographical framing in the main text — Tye's three decades developing tracking representationalism — is drawn from his published corpus (esp. *Ten Problems of Consciousness*, MIT Press, 1995, and *Consciousness, Color, and Content*, MIT Press, 2000); it is included to mark the stakes, not as biographical decoration. Tye opens the 2024 essay by citing John Perry's quip that if you study consciousness long enough you either become a panpsychist or go into administration — Tye's wry way of registering that he regards the conversion as a surprise to himself.

**6.** Tye, "How I Learned to Stop Worrying and Love Panpsychism," 14–20. The sharpness argument: phenomenal consciousness admits of no borderline cases (we cannot imagine a state for which there is no fact of the matter whether there is something it is like to be in it), and we know this a priori; every physical/functional candidate for identification with consciousness is vague (admits borderline cases); a sharp property cannot be identical to a vague one; brute emergence of a "new, non-physical, and sharp" property is unacceptable; therefore consciousness is a fundamental intrinsic feature of basic micro-entities. In n. 4, Tye identifies the no-borderline-cases premise as shared by Jonathan A. Simon, "Vagueness and Zombies: Why 'Phenomenally Conscious' Has No Borderline Cases," *Philosophical Studies* 174, no. 8 (2017): 2105–2123.

**7.** Michael V. Antony, "Vagueness and the Metaphysics of Consciousness," *Philosophical Studies* 128, no. 3 (2006): 515–538, conditions his metaphysical conclusions on both the sharpness and correctness of our present concept. His "Are Our Concepts CONSCIOUS STATE and CONSCIOUS CREATURE Vague?," *Erkenntnis* 68, no. 2 (2008): 239–263, argues for conceptual sharpness. Tye, "How I Learned to Stop Worrying and Love Panpsychism," 16 n. 4, explicitly describes Antony's 2006 view as allowing consciousness to be vague even though our current concept is sharp. The recognitional/demonstrative explanation in the main text is the book's proposal, not an account attributed to Antony. Tye's 2024 note directs readers to *Vagueness and the Evolution of Consciousness* (2021; cited n. 5) for his criticisms; the essay itself states that "it is part of the nature or essence of phenomenal consciousness that it lacks such cases" and that we "know a priori" that consciousness is sharp (17). Appendix B.10 develops the dispute over that claimed insight. For an independent defense of borderline consciousness and the argument that our inability to conceive such cases does not settle their metaphysical possibility, see Eric Schwitzgebel, "Borderline Consciousness, When It's Neither Determinately True nor Determinately False That Experience Is Present," *Philosophical Studies* 180 (2023): 3415–3439.

**8.** William James, *The Principles of Psychology* (New York: Henry Holt, 1890), vol. 1, ch. 6 ("The Mind-Stuff Theory"), 158–160. James writes on p. 158, "Each remains, in the sum, what it always

was"; he then turns on p. 160 to the hundred-feelings case: "Take a hundred of them [feelings], shuffle them and pack them as close together as you can (whatever that may mean); still each remains the same feeling it always was, shut in its own skin, windowless, ignorant of what the other feelings are and mean. There would be a hundred-and-first feeling there, if, when a group or series of such feelings were set up, a consciousness belonging to the group as such should emerge. And this 101st feeling would be a totally new fact." James is arguing against the "mind-stuff" theory — the nineteenth-century ancestor of constitutive panpsychism — and the passage is the locus classicus of the combination problem.

**9.** David Chalmers, "The Combination Problem for Panpsychism," in *Panpsychism: Contemporary Perspectives*, ed. Godehard Brüntrup and Ludwig Jaskolla (Oxford: Oxford University Press, 2017), 179–214. Chalmers distinguishes the *subject-combination*, *quality-combination*, and *structure-combination* problems and concludes that they remain serious though not obviously fatal. Philip Goff, "The Phenomenal Bonding Solution to the Combination Problem," in the same volume, 283–302, grants the isolation of subjects absent a further relation, then proposes a phenomenal bond that could necessitate a macro-subject; he argues that its opacity fits our epistemic position and canvasses identification with the spatial relation. Chalmers's sec. 6.3 examines phenomenal bonding and co-consciousness while stressing the further question of what the bond is. The original objection is James's (n. 8); the name "combination problem" is commonly credited to William Seager, "Consciousness, Information and Panpsychism," *Journal of Consciousness Studies* 2, no. 3 (1995): 272–288. The top-down inversion — cosmopsychism — trades combination for *decombination*. It develops from Jonathan Schaffer's priority monism, "Monism: The Priority of the Whole," *Philosophical Review* 119, no. 1 (2010): 31–76. For hybrid cosmopsychism's strongly emergent local subjects, qualia transfer, and Localization and Thinning Principles, see Philip Goff, "How Exactly Does Panpsychism Help Explain Consciousness?" *Journal of Consciousness Studies* 31, nos. 3–4 (2024): 56–82, esp. sec. III.

**10.** Tye's official mechanism appears in "How I Learned to Stop Worrying and Love Panpsychism," 20–24. A conscious macrostate, on his account, has consciousness* plus representational content; consciousness* does not suffice for consciousness by itself. Global-workspace theory then specifies when consciousness* transfers from microstates to a complex whole: once the relevant workspace arrangement is definitely in place and its contents are widely broadcast. Tye's dancers and words analogies motivate the transfer but do not explain why a subject-involving phenomenal feature should pass to the whole, or why a confessedly vague broadcasting organization supplies a sharp bearer-change. The charge in the text therefore concerns the added fundamental determinable and primitive transfer, not formal idleness: if sharpness independently requires consciousness*, it has a genuine theoretical job. For the general structural-versus-categorical cost debate, see Amy Kind, "Pessimism about Russellian Monism," in *Consciousness in the Physical World*, ed. Alter and Nagasawa (Oxford: Oxford University Press, 2015), 401–421. The book's conditional comparison locates phenomenal character in suitably poised, embodied, world-directed representational content and pays instead for a strong a posteriori identity; the verdict turns on which primitive package earns the better explanation.

**11.** Constitutive panqualityism treats fundamental qualities as protophenomenal quiddities without making them experiences of microsubjects; see David J. Chalmers, "Panpsychism and Panprotopsychism," in *Consciousness in the Physical World*, ed. Torin Alter and Yujin Nagasawa (Oxford: Oxford University Press, 2015), 246–276. Chalmers also locates John Heil's powerful-quality ontology within Russellian identity theory: dispositional and categorical descriptions concern one property, with either an a priori or a posteriori connection; see Heil, *The Universe as We Find It* (Oxford: Clarendon Press, 2012). Philip Goff's "Real Acquaintance and Physicalism," in *Phenomenal Qualities: Sense, Perception, and Consciousness*, ed. Paul Coates and Sam Coleman (Oxford: Oxford University Press, 2015), 121–143, states the strongest revelation claim. Marcelino Botin argues that introspection does not reveal phenomenal character as absolutely intrinsic in "Russellian Physicalists Get Our Phenomenal Concepts Wrong," *Philosophical Studies* 180, no. 7 (2023): 1829–1848, https://doi.org/10.1007/s11098-023-01955-1. Christopher Devlin Brown replies that phenomenal concepts may reveal nonfundamental physical categorical properties in "Russellian Physicalism, Phenomenal Concepts, and Revelation," *Philosophia* 51, no. 5 (2023): 2715–2722, https://doi.org/10.1007/s11406-023-00697-y. That reply supports local disclosure without establishing qualitative inheritance at the fundamental level.

## Chapter 11: The Gap That Isn't There

> *"There is its qualitative character, how it feels; and what is left unexplained by the discovery of C-fiber firing is why pain should feel the way it does!"*
>
> — Joseph Levine, "Materialism and Qualia: The Explanatory Gap"

CHAPTER OVERVIEW

The hard problem of consciousness can feel like a hole in nature: why should any physical process feel like anything at all? I argue that the gap belongs to how we know, not to what exists. We reach one fact by two routes: through a public account of brains and behavior, and through acquaintance with experience. If felt character consists in content of the right kind, no second felt fact waits for physics to produce. I defend that identity as a general thesis, but the book's stronger support comes from perceptual cases; pain's badness and mood remain unfinished applications. The general deflation inherits that limit: an identity can retire a demand for a bridge, but it cannot excuse the work still needed to support it.

### 11.1 Coffee and the Remaining Question

Take Chapter 2's coffee. Suppose neuroscience has mapped every receptor, nerve volley, cortical event, memory, and report. You hold the complete third-person description, with nothing physical omitted. Read it back. It still seems silent on why the processing feels like anything. It names the machinery behind the warmth, faint bitterness, and the way the roast announces itself while appearing to skip their delivery.

"I still feel a gap" is a reasonable response. The coffee has a taste; a description of your nervous system does not. Nothing in the identity claim asks you to deny that difference. But a description need not taste of coffee to explain tasting, any more than a description of digestion must digest. The serious question concerns what the explanation leaves out, not which activity the description fails to perform.

Two questions now need to stay apart. **What warrants identifying the feel with this physical, world-presenting activity?** Chapter 6 offered a comparative answer. Blur and attention support its account of perceptual contrast, but do not select its historical width over a structural identity. **If that identification holds, what further feel must the activity produce?** None: that demand counts the same occurrence twice. Answering the second question does not strengthen the evidence for the first.

I defend an *epistemic* gap, a gap in how we know, rather than an *ontological* gap in what exists. Having the experience gives us a way of recognizing it that reading the complete account need not give us. The warmth and bitterness remain; their physical explanation need not recreate them in its reader. That distinction alone does not answer Levine's harder demand for understanding.

We need to see why the two ways of knowing could remain apart even if Chapter 6′s identification succeeds.

### 11.2 What Is Really Left Over

Chapter 9 distinguished Joseph Levine's *explanatory gap* from David Chalmers's modal route to a gap in nature.[1] Levine asks why a physical description does not make felt character intelligible, even if some psychophysical identity holds. The liquidity of water follows, once understood, from the behavior of its molecules. Why this physical state feels like this does not seem to follow in the same way. Levine treats that failure as a problem for materialism, not as proof that materialism is false. It also threatens our warrant for deciding *which* identity holds: what physical or functional similarity would justify saying that a differently built creature shares our pain?[2] Chalmers adds a modal bridge. If we can coherently imagine, even after ideal reflection, a world with all the physical facts and no phenomenal fact, he argues that the scenario describes a genuine possibility. A possible physical duplicate without experience would refute physicalism.[3]

Two sentences easy to run together drive the debate:

> *Levine's sentence:* A complete physical description does not make it intelligible why the relevant state feels the way it does.
>
> *Chalmers's modal conclusion:* If ideal reflection leaves a physical duplicate without the feel coherently imaginable as a way the world might actually be, that scenario describes a possible world. The physical facts therefore do not necessitate the phenomenal facts.

I accept Levine's intelligibility complaint. A type-B physicalist, who expects an identity not reached by conceptual analysis alone, can accept it too; not every physicalist does. Chalmers's conclusion requires the modal bridge, and this book refuses that bridge. The move from "not made intelligible by" through ideal conceivability to "not necessitated by" and "not identical to" supplies the ontological threat.[4] Refuse that move and the explanatory gap no longer establishes a split in nature. That is my conclusion, not Levine's. The gap can still bear on our warrant for an identity. Chapter 6′s comparisons support ordinary perceptual contrasts more strongly than historical width or affective character. Calling the gap epistemic does not discharge those debts.

### 11.3 The Two Concepts of One Fact

Morning star and evening star name Venus. Ancient skywatchers tracked it at dawn and dusk by two routes without knowing the routes met. Their discovery taught them something genuine but supplied no *mechanism* connecting one star to another. No one asked how Venus manages to be Venus. No bridge runs between them because there was only one planet, caught in two ways.[5]

And yet, even after you know the identity, the two ways of catching sight do not collapse into one. You can still think of Venus *as the morning star* and think of it *as the evening star*, and these remain different thoughts about the same planet. The identity reports a fact about the planet. The persisting two-ness, a fact about your concepts. The first does not abolish the second, and it would be a confusion to expect it to.

Now bring that structure home. Chapter 6 argued that felt character consists in a world-directed representational activity of the right embodied kind. Suppose that is true. We have one fact and two ways of reaching it. From the outside, we use *physical and functional concepts* — concepts of firing patterns, representational role, and the cascade from receptor to report. From the inside, by undergoing it, we use a different kind of concept altogether.

Much of the literature calls the resulting ways of thinking *phenomenal concepts*, though the label carries theories this book need not inherit.[6] What matters here is the underlying capacity. You acquire it by having the experience: *what red is like*, *what that ache is like*, formed in the undergoing rather than assembled from a description. For a person with our architecture, a lecture on wavelengths, cone types, and cortical maps does not supply the sight-restored adult's recognitional *this*. The relevant experience, or one close enough to it, supplies that anchor. Mary knew every physical fact about red before her release, then gained an acquaintance-grounded way of representing from the inside a property she already knew from the outside.[7] Her thought has an indexical component (*this state, now*) but is not exhausted by self-location. The demonstrative anchors the present token; the recognitional capacity lets her classify later occasions. Extending that classification to other perceivers requires evidence about their organization and circumstances. Acquaintance with her own state does not certify theirs. Pain, taste, bodily sensation, and emotion invite the same account of acquisition without guaranteeing the same success in every application. Mary's file on red was complete. She lacked a way of reading a page she already had.

Acquiring a concept and connecting it to another concept are different achievements. Once Mary has both, empirical evidence can warrant the identity between what she recognizes and what neuroscience describes. The two routes need not remain inferentially disconnected in that sense. The narrower claim is that the recognitional concept does not itself provide an analysis in physical terms, nor does the physical concept by itself install the recognitional capacity. The house account models the first route through a demonstrative and learned recognition, the second through causes, organization, and effects. No rule of equivalence comes built into that distinction. This explains why acquiring the first-person concept need not make the identity transparent. It does not prove that every possible knower lacks an a priori route; Mary* will test that stronger claim below.

Call the house version the *acquaintance-recognition account*. It sits within the debate over the phenomenal-concept strategy, the contemporary physicalist reply to Levine developed by Brian Loar, David Papineau, Katalin Balog, and others. Tye later rejected the need for a special class of phenomenal concepts while retaining a materialist role for acquaintance.[8] The book needs no theorist's proprietary machinery. It needs the narrower structure: undergoing can supply a recognitional capacity that public description alone does not supply, without thereby making the identity conceptually transparent. *Two routes to a fact need not reach two properties.* Whether these routes reach one physical fact depends on the identity's independent warrant.

The account gives Mary a genuine cognitive gain. An occurring state anchors *this taste*; later recognition can reuse a capacity she previously lacked. She can now entertain a thought about how the taste presents itself, rather than only about its public description. That difference explains

the new thought's significance without guaranteeing its truth or justification. A mistaken classification can use the same capacity. Whether a judgment counts as knowledge depends on what anchors it, how reliably recognition works, and what defeats or supports it in the circumstances. Those questions require more than concept acquisition. Section 11.7 therefore limits the strategy: it explains acquisition and a possible source of conceptual asymmetry, not the whole truth-involving epistemic gap Chalmers challenges.

Section 6.4.3 retains the wide identity on Inverted Earth. It predicts that character can change without detection because "the standards of comparison move with it." That extension has less support than the broader intentionalist account. A structural rival can preserve memory's present organization and character without adding mental paint. Acquaintance supplies no independent check on the predicted drift.

Nor does acquaintance buy an infallible gauge back. It works in the present tense: *this*, while the coffee does its work. The occurring state anchors the token's reference. A subject can misdescribe or misclassify the state while still picking it out then and there. On the wide identity, character follows the content the state then carries. If content shifts during the episode, character shifts with it. Local anchoring follows the occurrence rather than preserving an inner sample. Memory must do the work of comparing times, and its own content has a history. Present acquaintance gives no verdict on whether that history preserved the earlier character.

The gap lives in the knower, and its public life gives another clue. Gap-*talk* (the conviction, the protest, the journal literature, this very paragraph) consists in physical events in physical beings. The completeness of the physical, established in Chapter 9, guarantees each a sufficient physical cause. Whatever consciousness turns out to be, judgments about it sit squarely in the causal order, with biographies a finished neuroscience could write line by line.

Chalmers did not wait to be pressed on this; he raised it against himself, named it (the *paradox of phenomenal judgment*) and accepted it: on his own view, the processes that produce your conviction that consciousness outruns the physical would run identically whether or not it did.[9] His zombie twin, remember, is physically identical; so the twin writes the same dissertation on the hard problem, mounts the same defenses of the gap, feels (no, *registers*) the same certainty, all of it produced by the same physical causes, and nearly all of it, by hypothesis, wrong about itself. Chalmers's reply is that the evidence never was the conviction; it is the experience itself (present in him, absent in his twin) and a phenomenal belief is justified by what acquaints, not by what causes. Allow him the structure; it is not absurd, and he has defended it with care. First-person justification need not be third-person auditable. The comparative cost is explanatory bifurcation: the physical process explains every public and functional feature of the judgment, while the experience receives a justificatory role that leaves that process unchanged. The book speaks of acquaintance too, but claims no greater introspective authority for it: its acquaintance is a physical relation a being stands in to its own world-directed activity, squarely inside the causal story of what it says. The epiphenomenalist's acquaintance, under this package, leaves no trace in the story of the judging. That does not refute the appeal to acquaintance. It leaves the judgment and the fact offered as its evidence in separate explanatory histories. Identity avoids that particular separation, though it must still explain how conscious activity produces the relevant judgments.

Chalmers turned the observation into a research program. His *meta-problem of consciousness* asks why creatures like us judge there to be a hard problem at all; by his own classification it belongs with the tractable problems, answerable in physical and functional terms.[10] By itself, such an explanation supports neither identity nor dualism. Chalmers expects it to leave realism about consciousness standing, and an unconverted reader has not granted the identity that would decide the matter. The comparative pressure concerns the epiphenomenalist package just considered. There the processes producing certainty would run identically without experience. On the identity account, conscious activity belongs within physical causation; the task is to explain how it generates judgments about its character. Identity makes that integration possible without a second justificatory history. It does not complete the explanation or exempt physicalism from the meta-problem. Chalmers also considers nonreductive accounts on which consciousness helps realize the judgment-producing processes. The comparison here does not dispose of those alternatives. Objection 1 presses the epistemic limit.

### 11.4 Why an Identity Needs No Bridge

When two concepts present one fact, the demand for a *bridge* between them miscounts. No two stars wait to be joined, only one planet doubly named. Ask why water *is* $H_2O$, why these molecules should constitute this liquid rather than cause or correlate with it, and a warranted identity leaves no further production step. It is *explanatorily terminal* in that restricted sense: water's properties get explained through $H_2O$'s behavior without $H_2O$ then having to manufacture water as something else. Evidence for the identification and explanations of the molecular organization remain proper questions. That is where the spade turns, not where inquiry stops.[11]

Applied to consciousness, the hard problem changes character. The inner theater posed a connection problem: *here* physical processing, third-personal and objective; *there* felt experience, first-personal and subjective. But a connection problem between two things cannot survive an identity that shows there are not two.

A dualist can object that the phenomenal case differs. With water, an independent third-personal grip on the liquid, its rolling, freezing, and dissolving behavior, follows from the molecular story. Felt character is reached first through acquaintance; where is the third-personal explanandum the identity closes over? The analogy cannot answer by itself. It works only if Chapter 6′s independent case warrants identifying felt character with representational activity. Section 11.3 then offers an account of how their conceptual routes could remain asymmetric. Refuse the identity and the disagreement returns to Chapter 6. Venus illustrates the no-additional-product point; it does not establish the psychophysical identity or explain every feature of our knowledge.

On the identity claim, felt experience consists in world-directed representational activity: one event described from outside and lived from inside as seeing red or tasting coffee. The demand for an additional product then has the grammar of "why does Venus give rise to Venus?" But "why accept this identity?" remains a proper question. The analogy cannot dismiss it. A critic who asks for better evidence, or proposes necessary grounding instead, has not miscounted merely by declining our answer. The connection problem retires only as far as the identification earns its warrant.

A dualist has one charge left, and it is the sharpest available: that an identity of this kind could hold only as a *strong necessity*.[12] Section 11.7 takes it up.

What remains (and I do not deny that something remains) is not that question but a tamer successor, tamer in what it threatens, not in how hard it presses. With the identity in place you can still ask why nature contains world-directed activity at all, how physical organization realizes it, why its contents differ as they do, and how strongly a given state realizes the distributed pattern of recurrence, availability, correction, and control.[13] Those are real questions, and the book does not claim to make them vanish. But they concern one physical world's processes and distribution, with no second realm hovering behind them and no bridge left to build. The menacing hard problem asks why such a process *additionally* produces a feel. Conditional on Chapter 6′s identity, that demand double-counts one process entered twice.

Identity does not make the felt residue disappear. Paired with the acquaintance-recognition account, it shows how a difference in access could persist without requiring another property. Our present theory remains incomplete. Physical realization, inferential relations between concepts, and the warrant for phenomenal knowledge still need explanation. The gap that isn't there, *if the identity holds*, is the ontological gap. Levine's explanatory challenge remains, including its pressure on which identity we should accept.

### 11.4.1 Where a Conscious State Enters the Story

Deflating the ontological gap does not license leaving state consciousness blank. Chapter 6 makes a positive, defeasible proposal: a content-bearing state becomes conscious for a subject through its sustained, selective availability across that subject's integrated economy. Under suitable conditions it can guide attention, working memory, inference, planning, report where possible, and flexible action. Feedback can preserve, revise, or displace it. This is a proposed account of *phenomenal* state consciousness, not just a definition of access. Ned Block distinguishes being experienced from being available for reasoning and action. Their coincidence therefore needs defense, not a change of labels.[13a] Breadth, reuse, stability, and two-way correction specify the proposed connection. No single behavioral effect suffices.

Dennett's Multiple Drafts model supplies the useful contrast with the theater. Processing occurs in parallel, revisions happen at many sites, and no last editor presents a finished result to an audience. Some contents nonetheless gain "fame in the brain": wider and more durable effects on what the system remembers, says, reasons about, and does. That picture gives *poising* a tractable meaning. Poise names a content's place in a distributed competition for access and control. Recurrence offers one way to sustain that place; a particular feedback loop is not the definition of consciousness.

Consider a briefly presented word that alters one later response but cannot be retained, reconsidered, or used flexibly elsewhere. The proposal predicts unconscious processing, not a faint experience merely because something influenced behavior. A scene sustained across attention and flexible use can instead count as conscious even when no report is requested or remembered. Reach and stability come in degrees; that does not yet settle whether experience itself has a sharp boundary or what combination of capacities suffices at the margin. Global-workspace, recurrent-

processing, predictive, and higher-order theories offer competing accounts of realization. The proposal has not settled their thresholds.

Testing it requires keeping the evidence apart. Content-sensitive interventions can measure retention, cross-system use, and correction without defining consciousness by a verbal answer. Separately, sincere reports of presence or absence, discrimination, confidence, and contrasts between visible and masked presentation can provide fallible evidence about experience. Chapter 6 adopts that discipline, treating reports as evidence rather than accepting their metaphysical interpretation. A stable report of experience where the independently measured profile was absent would count against the proposal. Convergent evidence of absence despite a matched profile would press from the other side; silence alone would not establish it. No test supplies a theory-free guarantee of consciousness. The commitment is to let the two bodies of evidence disagree, not to classify every inconvenient case by the profile. Among states counted as conscious, content, perceptual mode, determinacy, and organization then address *what* the experience is like. That character identity faces its own matched-profile test.

### 11.5 Emergence and Identity

"Consciousness emerges from the brain" can name a perfectly legitimate research question. How do interacting cells acquire the organized capacities no isolated cell possesses? A starling flock's wheeling requires no extra bird above the flock, and an emergent pattern need introduce no new substance. *Weak emergence* names the surprising high-level regularities that arise from lower-level organization without requiring new fundamental ingredients. Complexity can make them difficult to predict or understand. It does not make them nonphysical.

*Strong emergence* adds a further claim: the higher-level phenomenon requires something not supplied by the physical base and its laws. Some versions propose irreducible downward causal powers; others add basic psychophysical laws without such powers. Chalmers's discussion distinguishes these possibilities.[13b] Nor should failure of *a priori* deduction alone settle the matter here. This book's type-B physicalism already accepts that failure while defending a necessary identity. Moving from what we cannot derive in thought to what nature must add would repeat the disputed inference.

The question, then, concerns what emerges *from what*. We still need to explain how brains develop and sustain conscious, world-directed organization. The identity claim offers no exemption from that work. What it rejects is a further production step from the complete conscious world-presenting profile to its feel. Phenomenal character consists in that profile; it does not arrive afterward. Calling the organized activity weakly emergent can coexist with this claim. Calling its character a further irreducible product contradicts it. The word cannot choose between those accounts for us.

### 11.6 What the Deflation Depends On

Everything in Section 11.4 depends on a real identity. The deflation stands *hostage to the identity claim*: felt character must consist in world-directed representational activity, not merely travel alongside it.[14] Ned Block pressed the point sharply. The acquaintance-recognition move cannot manufacture the identity; it can only explain why, *given* the identity, an explanatory gap persists.

Without the identity, this particular stopping argument is unavailable. A grounding rival would need to earn a different constitutive explanation.

"Given" does not mean *merely stipulated*. Chapter 6 gives a worked account of blur and attention, then makes two further choices. It chooses identity over grounding, and wide over structural character. Economy favors the first choice. Unity motivates the second, but these comparisons do not test it. A red-looking white tile shows the motive: one property attribution describes the look and specifies the error. A structural identity can also preserve the look while assigning wide reference a separate role in accuracy. That rival can deflate the hard problem too. Retiring the bridge demand gives no advantage peculiar to my wide identity. I retain it, while its historical claims inherit Chapter 6′s weaker support.

Section 11.3′s comparison of reports has the same limit. If identity holds, the experience and the judgment about it belong to one physical-causal history. An epiphenomenal account separates the experience's justificatory role from the physical production of the report. That is a comparative cost of the rival, not an independent observation that our identity holds; a residue-free physicalist grounding theory need incur no such separation. Neither the persistence of gap-talk nor our ability to explain it settles that closer contest.

The line between identity and correlation falls most instructively on John Searle. He holds that consciousness is caused by and realized in the brain, an ordinary biological feature of the organism, as natural as digestion.[15] His familiar "simple solution" to the mind-body problem blames an inherited Cartesian vocabulary whose categories, *mental* and *physical*, were built to exclude each other. Drop the vocabulary and the problem dissolves. That instinct lies close to this chapter's own: both diagnoses locate the mystery in how we frame one natural world.

Yet Searle's official account relates neural and conscious by *causation* rather than reductive identity. Lower-level neural processes cause a system-level conscious feature while the whole remains one biological system described at different levels. This introduces no second substance or realm. The narrower problem suffices. So long as the conscious feature remains ontologically irreducible to its neural realization, "why does this organization have *that* felt character?" stays open in the form the identity claim retires. Dissolving-by-vocabulary requires strict identity as the terminal point. Searle declines that identity, so his causal-realization relation cannot perform this chapter's stopping move.

Biological naturalism can therefore reject dualism and keep consciousness wholly biological. What it cannot do is use identity to retire a bridge it declines to posit. A causal account may explain when consciousness occurs, what sustains it, and what it does while leaving the explanatory demand intact. The book follows Searle on perception but departs from him here: only the identity step his framework refuses licenses this deflation.[16] If that identity falls, the deflation falls with it.

The scope of that support also matters. Part Three explains how history and use can fix what a state gets right or wrong.[17] Its worked perceptual cases support the identity thesis more securely than its affective proposals do. A blurred edge gives us a contrast to test against the visual profile. An aching ankle poses further questions: what fixes the content of badness, and what warrants wanting this experience itself to end? Chapters 7 and 15 do not answer those questions. The

claim remains universal: phenomenal character consists in content of the right kind, one and the same thing under two descriptions. But the achieved explanation does not yet cover every kind of character that thesis identifies.

The general deflation therefore remains conditional on an identity whose support varies across cases. I defend an epistemic gap, not a split in nature. That does not license treating our unfinished account of pain as though it were merely a difficulty in grasping an identity already established. Nor does it license denying pain's badness. Content, character, state consciousness, and the subject still ask different questions. A test for conscious access cannot supply the missing evaluative content, and a theory of the subject cannot supply the grounds for relief.

## 11.7 Objections

### Objection 1: The phenomenal-concept move cannot explain our whole epistemic situation (Chalmers's dilemma)

Chalmers begins with a demand for neutrality. A physicalist may invoke some psychological feature, call it C, to explain why Mary learns, why zombies remain conceivable, and why the gap seems permanent. C cannot mean *phenomenal acquaintance*. That would put the disputed experience into the premise. It must describe the relevant architecture without deciding whether physicalism is true.

If C could vary while every physical fact stayed fixed, physicalism has merely moved the gap into C. If C could not vary, Zombie Mary has it too. Yet the zombie's corresponding belief may lack the truth or justification Mary gains. C can then explain their shared recognitional dispositions and gap-talk, but not their whole epistemic difference.[18]

The book takes the second horn with a limited ambition. Let C describe an occurring state anchoring a demonstrative and a capacity for later recognition. A physical duplicate shares that architecture. C can explain acquisition and shared gap-talk. It does not alone establish inferential isolation, successful reference, or warranted phenomenal knowledge. If Chapter 6′s identity holds, Mary's demonstrative can reach a physical truthmaker; the identity does not by itself explain why her belief qualifies as knowledge. Chalmers explicitly allows an augmented account of reference and justification, then reapplies his dilemma to it. Invoking identity cannot answer that challenge without borrowing the conclusion at issue. Taking this horn therefore does not independently undercut his inference from a truth-involving epistemic gap to an ontological gap. The book resists that inference through the modal disagreement in Chapter 9 and Objection 4, with Chapter 10 comparing the alternatives. Here the claim is narrower: concept acquisition can differ while the fact reached remains one, if the identity earns its warrant. A structural identity can share that answer. Neither our historical extension nor a complete account of phenomenal knowledge follows from it.

Philip Goff asks what first-person acquaintance actually reveals. If it reveals nothing, phenomenal thought becomes a rigid tag. If it reveals part of pain's nature, the type-B physicalist must explain how that disclosure can concern one wholly physical property without adding a second phenomenal aspect.[19]

Goff's inference needs more than the claim that acquaintance reveals something real. It needs the further claim that revealing part of a property's nature makes that aspect available under the concepts of a complete physical description. I deny that step. Acquaintance presents the coffee's bitterness or the body's painful disturbance without giving the taster criteria for identifying that presentation with a neuroscientific kind. Undergoing supplies the recognitional anchor; evidence must still connect it to the physical account. Disclosure of a character and disclosure of its place in a theory are different cognitive achievements. Calling the disclosed aspect physical, *if the identity holds*, does not make the latter achievement follow from the former.

Acquaintance lets the taster judge that this sip seems bitter rather than sweet, or locate an ache in the ankle rather than the hand. These judgments classify a presented character; they do more than attach a bare tag. They remain fallible and supply no analysis that identifies their referent under physical concepts. I grant this richer presentation and recognition while denying that it supplies Goff's stronger grasp of essential nature. If Goff's *translucency* requires that stronger grasp, I do not grant translucency. If it means only that how the state presents itself becomes available to the subject, I grant it but deny the claimed a priori consequence. The difference is not a second aspect hidden behind the first. One property can be presented without an analysis that lets us identify it under another vocabulary. This is the house answer to Goff, not a proof that the identity holds. He can reject its account of disclosure; acquaintance alone does not decide between us.

Helen Yetter-Chappell presses from the opposite direction. Her Mary* already has complete general physical knowledge and genuine phenomenal concepts. An unusual cognitive architecture lets her match recognitional information to physical descriptions automatically. Yetter-Chappell argues that the resulting knowledge of psychophysical identities can be *a priori*: if such a knower is possible, the ideal epistemic gap claimed by type-B physicalism disappears. Her target is type-B physicalism, not physicalism itself.

The distinction between acquiring a concept and justifying an identity now matters. Mary*'s phenomenal concepts are part of the starting conditions. The case does not show that physical description supplies them without acquaintance. Nor would experience needed to acquire them automatically make a later identity judgment a posteriori: enabling experience and evidence for a judgment are different. A sound reply cannot dismiss her merely for having different wiring.

What the case would defeat is an unrestricted claim that no possible knower could have a priori knowledge of the identity. I do not derive that claim from the acquisition account. For our acquaintance-grounded concepts, no a priori analysis follows merely from how we acquire them. An automatic match, considered by itself, does not establish the warrant for an identity; Yetter-Chappell's stronger claim requires that the matching deliver knowledge of the same phenomenal–physical proposition. If it does, the universal ideal-gap thesis must go. The narrower account survives: acquaintance can still enable the phenomenal concept, and our two routes can remain nontransparent. Chapter 9′s concession concerns ideal reflection with the concepts used in its zombie argument, not a proof about every possible architecture. I retain that concession and the dispute over its modal force. Mary* is a limit on what this chapter can establish about the type-B classification, not an objection answered by observing that she adds no second property.

**Objection 2: You have only relocated the gap; it still *feels* metaphysical**

The opening objection now returns with a sharper point:

> *I understand why reading about coffee need not give me its taste. But I already have the taste, and I understand your proposed identity. Why does the taste still seem to disclose something your relational description misses?*

Grant the datum in full: the red looks this way, the ache hurts this way, the coffee has this warmth and bitterness. The relational account preserves that felt reality. It disputes an interpretation of the datum — that introspection also reveals an intrinsic qualitative constituent, separate from the surface, disturbance, or warmth as presented. Calling that interpretation *inner paint* does not make the paint vanish; neither does feeling its pull establish it.

Chapter 5 taught a useful test. Attend as closely as you can to what remains after the presented red, the bodily disturbance, and the warmth have been specified. The exercise may leave a powerful sense of *this*, but it does not disclose a second feature with an independently describable profile. That result underdetermines the metaphysics. The intrinsic-quality theorist can say the elusive this-ness is precisely what acquaintance reveals; the intentionalist can say acquaintance has re-presented the same world-directed character under a recognitional concept. The argument cannot be settled by ordering the reader to look harder.

Introspection secures the felt character and its first-person presentation. It does not by itself show that character is intrinsic. The identity account explains the datum and the stubborn sense of residue through the split between acquaintance and description. A theory that adds a further property faces the costs already counted. A structural identity adds no such residue and can share this account of the gap. The dispute over history needs separate evidence. No phenomenon gets eliminated. What loses its automatic authority is the inference from *there is something it is like* to *therefore an intrinsic feature lies beyond every relational account.*

One strand of the objector's sensation survives every reply this section can give — not the inner paint, not the intrinsicness, but the bare wonder that there is phenomenal presentation at all, that any of this shows up. The chapter has not extinguished that wonder and does not try. But wonder is not yet the retired explanatory demand. Ask why this physical activity *additionally produces* a feel and the identity has double-counted; ask why nature contains beings in whom this activity occurs, how it is realized, why its contents differ, and where its distribution begins and ends, and real questions remain. Section 11.4 booked those questions among the open work. The bare astonishment belongs with them as an attitude toward the fact, not as evidence of a second fact left unexplained.

**Objection 3: The gap marks a failed framework, not a concept-pair (the later Nagel)**

Thomas Nagel, whose bat gave the field its permanent phrase for the problem, returned to it nearly forty years later with a far larger claim. *Mind and Cosmos* argues that the failure to explain consciousness is no local puzzle but a symptom: the materialist framework itself (physics,

chemistry, and the neo-Darwinian account of life) cannot in principle accommodate mind, and an adequate successor may need principles teleological in logical form.[20] On that view this chapter has misdiagnosed its patient. The gap does not run between two concepts of one fact; it runs between reality and an entire conceptual scheme, and no rearrangement of the scheme's existing pieces, the identity claim included, can close it. Waiting for the revolution is the honest posture, and deflation a premature consolation.

Nagel asks two questions that the reply must keep apart. What constitutes consciousness? How did beings with that constitution arise? He also challenges the identity directly: a causal-role account seems to leave subjective appearance out, while a necessary psychophysical connection may require concepts our present science lacks. His broader case is conditional on reduction failing; it is not simply a failure to notice the word *identity*.

Chapter 6 disputes the constitutive objection by identifying felt character with the world-presenting profile, not with behavioral effects that leave the appearance behind. Sections 11.3 and 11.4 explain the limited stopping point that follows. If that identity holds, no further feel must be produced after the complete conscious profile exists. Nagel can still reject the identification or demand a better explanation of its necessity. The disagreement over that demand remains live in Objection 4. Calling it an epistemic gap does not refute his argument.

The historical question remains even if the identity succeeds. Identity alone establishes neither an evolutionary trajectory nor the sufficiency of ordinary selection. Nagel's teleological proposal could address the formation of physical organization; it need not wait for a second substance to appear. The book's alternative is to explain world-presenting organization through selection, learning, and use. Part Three must show how that organization fixes content, and an evolutionary account must explain how organisms with it arose. A model of content fixing is not yet a complete evolutionary history. The identity removes the extra-product demand; it does not announce that those jobs have been done. Nagel has not thereby established the need for new teleological laws, but this chapter has not eliminated that proposal merely by adopting identity.

**Objection 4: The identity would be a strong necessity, and a strong necessity is dualism wearing a physicalist label**

This is the most precise objection a dualist can raise against the kind of physicalism this book holds. The book has called its view *type-B* physicalism, the a-posteriori sort, on which the psychophysical identity is true but not knowable by reflection alone, exactly like water and $H_2O$. Chalmers's reply is that the analogy breaks at the joint that matters. An ordinary a-posteriori identity, he argues, comes with a story that *dissolves* its appearance of contingency: water seems possibly-not-$H_2O$ only because we were fixing the reference of "water" by a contingent mode of presentation (the watery stuff around here), and once that mode is spelled out, the necessity follows and the air of contingency evaporates. The phenomenal identity has no such story. The felt character does not pick out its physical correlate by some further contingent feature we could peel away; the way it presents itself *is* its nature, with nothing behind it to do the reference-fixing. So the identity, if it holds, must hold as a *strong necessity* — a necessitation with no accompanying explanation of why it could not have been otherwise, and (Chalmers adds) with no analogue anywhere else in science, where every necessity earns its keep with a story. A brute necessitation

with no story and no precedent, he says, is not physicalism vindicated but dualism wearing a physicalist label.

The book's reply concedes the strong-necessity label and then asks what the label proves. Acquisition supplies one relevant asymmetry. A perfect account of facial geometry without perceptual exposure to Gordon does not hand you the ordinary ability to recognize him in a crowd. A perfect account of wine chemistry does not put the vintage on your tongue. These examples concern acquisition and non-inferential recognition only. They neither establish an ideal epistemic gap once the concepts are possessed nor supply an ordinary analogue of strong necessity. Chalmers can grant them and retain his modal objection. The phenomenal case must meet that objection on its own terms.

Chalmers can still insist that a necessary identity without an a priori route between the relevant concepts would stand alone in nature. Type-B physicalism accepts that revisionary cost; loyalty to physicalism cannot justify it. Chapter 6 supplied the comparative case, stronger for perceptual presentation than for historical width. Chapters 9 and 10 counted costs of its rivals. The view accepts strong necessity after that comparison, not solely as a rescue here.

The refusal concerns one exact principle of Chalmers's: ideal primary positive conceivability entails primary possibility. I grant the ideal zombie conceivability at issue in Chapter 9. The acquisition account does not prove that concession; nor does it refute the modal principle. Ideal reflection improves reasoning with the concepts given. Mary* asks whether a differently organized knower could have an a priori connection that our account does not supply. This chapter cannot turn the limits of our acquisition routes into a theorem excluding her. The local modal standoff remains: the adopted identity and Chalmers's bridge yield opposing verdicts about the granted scenario. I side with the identity and treat conceivability as defeasible evidence. That choice does not settle wide against structural character, complete the epistemology, or establish that every possible knower must lack an a priori route. The price remains strong necessity for the identity as framed in this dispute.

**Objection 5 (the tempting misreading): "So the zombie is impossible, and you've proved it here"**

A reader pleased by the deflation may want it to declare the philosophical zombie inconceivable. Resist the overreach. Chapter 9 grants ideal conceivability and challenges the inference to possibility. This chapter's narrower claim is that zombie imaginability may be the explanatory gap in modal clothing: the distance between acquaintance- and description-concepts projected onto possibility. The deflation need not win the zombie argument. Conceivability does not, without the disputed modal bridge, establish an ontological gap; its evidential weight remains part of the disagreement.

### 11.8 The Flip, Made Stable

The deflation can sound like a long subtraction, leaving a thinner world than the one with which we began. That gets the book backwards. Part Two has one last debt to settle.

Nothing has been taken from experience. Tomato red remains red, vivid, and yours; the ache still hurts; the coffee still tastes of exactly what it tastes of. Felt character is real because it *is* something

in the world — the world disclosing itself to a being built to be met by it. What the deflation removes is the theory that felt character is an inner exhibit requiring a non-physical ingredient or a fundamental dab of consciousness. Setting that theory down does not dim the red. It hands you back the tomato.

Return to the kitchen table. While the argument ran, the coffee cooled. A fresh sip brings less warmth and perhaps makes the bitterness more noticeable; the cup's weight stays in your hand while attention returns to the page. None of that requires inspecting a private exhibit before finding your way back to the kitchen. On the account defended here, the felt life consists in these changing, world-directed presentations, nothing over and above them. The explanation leaves the taste where we encountered it, in your tasting of the coffee.

The reversal concerns the place of consciousness in our picture. Instead of a private light sealed away from the world, we have an organism embedded in it: physical states present that world, recur through attention, memory, reasoning, and action, and belong to a persisting subject. The book rejects both an added nonphysical property and phenomenal character installed among nature's fundamentals.[21] Its positive claim remains open to the tests already stated. You need not lose your astonishment to understand it, or pretend the identification now feels obvious. At the kitchen table sits part of the physical world, meeting the rest of it, with something at stake in how the meeting goes.

The wonder changes address. It moves out of the private skull and back into the astonishing fact the theater hid: there is a part of the physical world, you, to which the rest of the physical world shows up.

Chapter 6 set the identity claims in a table, the book's own among them. The side this chapter has been refusing to walk is the straight line from mind to matter. A neuron still does not look like an ache, and dropping the sign on the cell, the role, or the program does not make the pictures match. With the identity in place, the demand for a further product double-counts one fact. The remaining side of the triangle waits in Part Three: how matter, through history and use, comes to present a world at all. Until that account arrives, the identity claim has named its content without yet showing how a physical state acquires one.

**Chapter Summary**

**The claim.** The hard problem needs deflation, not an added feature of nature. If the identity claim holds, description and acquaintance reach one physical fact by different routes, so the gap lies in our concepts rather than in the world.

**Where the argument stands.** A description need not give its reader an experience to explain it. Acquaintance grounds a recognitional way of thinking that public description alone does not supply. This accounts for acquisition and a possible source of conceptual asymmetry, not every feature of phenomenal knowledge. If identity holds, no additional feel must be produced. The argument neither establishes that identity nor defeats a necessary-grounding rival matching Chapter 6′s evidence. A structural identity can share the deflation. Historical width retains its weaker support.

**What's left open.** Inferential isolation and warranted phenomenal knowledge require more than an acquisition story. Mary* challenges a universal ideal-gap thesis, not the need for acquaintance to enable our concepts. The achieved explanation remains stronger for perception than for pain's badness, mood, and the grounds for relief. State consciousness, subject boundaries, evolutionary history, and the authority of ideal conceivability remain separate disputes.

**The hand-off.** Once the imagined bridge disappears, the world stands open again. The inquiry turns from how experience crosses into matter to how matter reaches beyond itself in meaning.

**Footnotes**

**1.** Joseph Levine, "Materialism and Qualia: The Explanatory Gap," *Pacific Philosophical Quarterly* 64 (1983): 354–361, esp. 354 and 357–61. Levine's mature statement is *Purple Haze: The Puzzle of Consciousness* (New York: Oxford University Press, 2001), esp. chs. 1 and 3, where he deepens the gap through the phenomenal-concept literature while keeping his materialist commitment explicit. The reading pressed here — that Levine catalogs an *explanatory* failing and stops short of Chalmers's metaphysical conclusion — is Levine's own; *Purple Haze* is at pains to distinguish the epistemic claim he defends from the ontological one he does not.

**2.** Levine's 1983 paper distinguishes the epistemological challenge from a proof that psychophysical identities are false, but treats it as serious pressure on materialism. Its closing argument, 359–61, asks what measure of physical or functional similarity would determine whether differently constituted creatures share our phenomenal states; the mere truth of some identity does not select the relevant kind. The present chapter accepts that constraint on generalization while denying that explanatory failure alone entails ontological difference. "Does not establish a split in nature" must not be read as "has no bearing on our warrant for the identity."

**3.** David Chalmers, "Facing Up to the Problem of Consciousness," *Journal of Consciousness Studies* 2, no. 3 (1995): 200–219, esp. 200–03, and *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press, 1996), esp. chs. 3–4. The conceivability argument from zombies — a being functionally and physically identical to a conscious person but phenomenally empty — runs in *The Conscious Mind* as a modal argument from conceivability to possibility to the non-supervenience of the phenomenal on the physical. This chapter engages only the inference *from the explanatory gap to an ontological gap*; the modal argument proper, which turns on the relation between conceivability and possibility, is the business of Chapter 9.

**4.** The progression — "not made intelligible by" → ideally conceivable without the phenomenal → "not necessitated by" → "not identical to" — distinguishes the explanatory datum from the modal and ontological conclusions. The book grants the datum and, in Chapter 9, the relevant ideal conceivability, while refusing the conceivability–possibility bridge. This is not a claim that ordinary a posteriori identities already resolve the phenomenal case: Chalmers's strong-necessity objection, addressed in Objection 4 and n. 12, presses that disanalogy. The type-B position bears the burden of its modal refusal and the independent case for identity.

**5.** The morning star / evening star case is the textbook illustration of an *a posteriori* identity, owed to Gottlob Frege's distinction between sense (*Sinn*) and reference (*Bedeutung*) — "Über Sinn und Bedeutung," *Zeitschrift für Philosophie und philosophische Kritik* 100 (1892): 25–50, esp. 25–27. Two

names (or concepts) can share a reference while differing in sense; the identity is informative because the senses differ, and it requires no "mechanism" connecting the referents because there is only one referent. Saul Kripke, *Naming and Necessity* (Cambridge, MA: Harvard University Press, 1980), esp. 97–105, 128–29, and 146–55, argues such identities are necessary though knowable only *a posteriori* — the model the phenomenal-physical identity claim follows, with the wrinkle (Kripke's own challenge) that the phenomenal case seems to resist the "appearance/reality" explaining-away Kripke uses elsewhere. The acquaintance-recognition account of Section 11.3 models a limited conceptual asymmetry; it does not by itself answer Kripke's stronger appearance/reality or modal challenge. The book's response depends on its separately defended identity and modal commitments. The distinct *modal* form of Kripke's challenge — that a genuine identity must be necessary, yet the phenomenal identity seems contingent — belongs with the conceivability arguments and is met in Chapter 9; the present chapter's point is only that a persisting explanatory gap does not by itself show the identity to be false. Its bearing on the identity's warrant remains acknowledged.

**6.** The literature uses "phenomenal concept" for the way of thinking about an experience that one acquires by undergoing it (or by imaginatively deploying a stored capacity derived from undergoing it). The classic statement is Brian Loar, "Phenomenal States," in *Philosophical Perspectives* 4 (1990): 81–108, esp. 87–89, revised in *The Nature of Consciousness*, ed. Ned Block, Owen Flanagan, and Güven Güzeldere (Cambridge, MA: MIT Press, 1997), 597–616. For the survey and critical taxonomy, see *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, ed. Torin Alter and Sven Walter (New York: Oxford University Press, 2007). For a map of the whole literature, Andreas Elpidorou, "Phenomenal Concepts," in *Oxford Bibliographies in Philosophy*, ed. Duncan Pritchard (New York: Oxford University Press, 2015), last modified January 15, 2015, reviewed May 11, 2023, doi:10.1093/obo/9780195396577-0254.

**7.** Chapter 8 defends the house mode-of-presentation reading: Mary gains acquaintance-grounded recognitional access and can entertain new true thoughts without thereby requiring a nonphysical truthmaker. Frank Jackson's original argument is "Epiphenomenal Qualia," *Philosophical Quarterly* 32 (1982): 127–136, esp. 130, and "What Mary Didn't Know," *Journal of Philosophy* 83 (1986): 291–295, esp. 291–94. His later physicalism differs from the house reply. In "Mind and Illusion," in *Minds and Persons*, ed. Anthony O'Hear (Cambridge: Cambridge University Press, 2003), 251–272, the section "The response that draws on a posteriori necessity" expressly rejects that response to the knowledge argument. In "Back to Mary on her release," Jackson instead reaches an ability account through representationalism: Mary's new representational state supports recognition, imagination, and memory. He explicitly agrees with Nemirow and Lewis and rejects the proposed new propositional knowledge that seeing red is like this. The book shares his representationalist and physicalist commitments without attributing its more concessive type-B account to him.

**8.** Brian Loar, "Phenomenal States" (cited n. 6); Katalin Balog, "In Defense of the Phenomenal Concept Strategy," *Philosophy and Phenomenological Research* 84, no. 1 (2012): 1–23, which defends a constitutional account and argues that the a priori dialectic does not compel either side to surrender; David Papineau, "Mind the Gap," *Philosophical Perspectives* 12 (1998): 373–388, and *Thinking About Consciousness* (Oxford: Clarendon Press, 2002), esp. ch. 2; the related "antipathetic

fallacy" diagnosis — that we mistake the limits of our imaginative-sympathetic grip on a brain state for a metaphysical limit — appears in Papineau's "Physicalism, Consciousness and the Antipathetic Fallacy," *Australasian Journal of Philosophy* 71, no. 2 (1993): 169–183, *Philosophical Naturalism* (Oxford: Blackwell, 1993), and "The Antipathetic Fallacy and the Boundaries of Consciousness," in *Conscious Experience*, ed. Thomas Metzinger (Paderborn: Schöningh, 1995), and is restated in the 2002 book; Michael Tye, *Consciousness, Color, and Content* (Cambridge, MA: MIT Press, 2000), and "Phenomenal Consciousness: The Explanatory Gap as a Cognitive Illusion," *Mind* 108 (1999): 705–725, where Tye argues that the gap is a cognitive illusion generated by the special features of phenomenal concepts rather than a window onto a metaphysical fact — a strategy Tye himself later abandoned (*Consciousness Revisited: Materialism without Phenomenal Concepts* [Cambridge, MA: MIT Press, 2009]) on his way to the conversion Chapter 10 examines. Daniel Stoljar introduced the label "phenomenal concept strategy" in "Physicalism and Phenomenal Concepts," *Mind & Language* 20, no. 5 (2005): 469–494; David J. Chalmers systematized and criticized the strategy under that label in "Phenomenal Concepts and the Explanatory Gap," in *Phenomenal Concepts and Phenomenal Knowledge*, ed. Torin Alter and Sven Walter (New York: Oxford University Press, 2007), 167–194 — the dilemma the text meets as Objection 1 below (see Appendix B.8).

**9.** David Chalmers, *The Conscious Mind* (cited n. 3), ch. 5, "The Paradox of Phenomenal Judgment." Chalmers grants that, given the causal closure of the physical, experience plays no role in the physical explanation of our judgments *about* experience — the zombie twin "writing chapter after chapter on the mysteries of consciousness" is his own image — and accepts the consequence, holding that phenomenal beliefs are justified by immediate acquaintance with experience rather than by their causal ancestry. He credits earlier statements of the paradox to Avshalom Elitzur and Roger Shepard. Chapter 9 met the same fact as the price of epiphenomenalism; here it returns as evidence about the intuition itself.

**10.** David Chalmers, "The Meta-Problem of Consciousness," *Journal of Consciousness Studies* 25, no. 9–10 (2018): 6–61, esp. 6–10 and sec. 5. The meta-problem is the problem of explaining, in physical and functional terms, why we judge that there is a hard problem — on Chalmers's own classification "strictly speaking one of the easy problems," since judgments and reports are behavioral phenomena. The paper canvasses the debunking arguments such an explanation would license; Chalmers argues each can be blocked, while conceding that a residual discomfort — a sense of fortunate coincidence between the judgments and the facts — survives the blocking. The deployment in the text takes that conceded residue and adds the closure premise Chapter 9 established, confining the causal comparison to the epiphenomenalist package. Section 4 distinguishes correlationism from realizationism; the latter allows consciousness to help realize the processes producing phenomenal intuitions. The program also challenges physicalist theories to explain those intuitions rather than identify their purported source without a mechanism. The program remains open, and Chalmers pursues it expecting realism about consciousness to be vindicated. The wing of the program that couples a meta-problem solution to a denial of the datum — illusionism — is engaged in Appendix B.16.

**11.** The point that *a posteriori* identities are explanatorily terminal — that one does not, and need not, explain why water *is* $H_2O$ over and above explaining water's properties by $H_2O$'s behavior — is developed by Ned Block and Robert Stalnaker, "Conceptual Analysis, Dualism, and the Explanatory Gap," *Philosophical Review* 108, no. 1 (1999): 1–46, esp. secs. 4–5, against the demand (pressed by Chalmers and Jackson) that physicalism must underwrite *a priori* entailments from microphysical to macro truths. Block and Stalnaker argue that reductive explanation does not in general require *a priori* entailment and that psychophysical identities can be explanatorily basic — exactly the structural claim Section 11.4 relies on. The "where the spade turns" image is Wittgenstein's (*Philosophical Investigations* sec. 217): explanation reaches bedrock, and the demand for a further explanation past the identity is the demand the metaphor is meant to retire.

**12.** The "strong necessity" objection to type-B (a-posteriori) physicalism is David Chalmers's, developed through his two-dimensional semantic framework: see "Consciousness and Its Place in Nature," in *Philosophy of Mind: Classical and Contemporary Readings*, ed. David Chalmers (New York: Oxford University Press, 2002), 247–272, esp. secs. 5–6; "Does Conceivability Entail Possibility?", in *Conceivability and Possibility*, ed. Tamar Szabó Gendler and John Hawthorne (New York: Oxford University Press, 2002), 145–200, esp. secs. 1–3 and 5; and "The Two-Dimensional Argument Against Materialism," in *The Character of Consciousness* (New York: Oxford University Press, 2010), 141–206, esp. secs. 7–10. An ordinary a-posteriori necessity such as water = $H_2O$ is weak in Chalmers's technical sense: the primary conceivability of watery XYZ reaches a possible world considered as actual, though actual water remains $H_2O$ in every counterfactual world. A type-B psychophysical identity would instead require a strong necessity, necessary in both dimensions, with no contingent reference-fixer to absorb the primary zombie scenario. The text accepts the classification and rejects the entailment from ideal primary positive conceivability to primary possibility. The acquaintance/description duality offers a model of conceptual asymmetry; acquisition conditions alone do not establish ideal conceivability or exclude every possible a priori knower. The book grants the relevant ideal conceivability in Chapter 9 while rejecting the modal principle. Neither that grant nor its rejection is proved by the acquisition model. For type-B replies, see Block and Stalnaker (cited n. 11), Brian Loar, "Phenomenal States" (cited Chapter 8, n. 14), and Stephen Yablo, "Coulda, Woulda, Shoulda," in *Conceivability and Possibility*, 441–492.

**13.** This is the precise sense in which the book's verdict is *reorganizes rather than dissolves* (the formula handed forward earlier in the chapter). The inner-theater version of the hard problem — how to connect two categorially alien things — is dissolved, because the identity shows there were never two things to connect. What survives are intraphysical questions about existence, realization, contrasts, boundaries, and distribution. The bare wonder that nature contains phenomenal presentation can remain, but it is not the further explanatory demand, "why does this process additionally produce a feel?" On the identity, that demand has counted the process twice.

**13a.** Daniel C. Dennett develops the Multiple Drafts model against the Cartesian Theater in *Consciousness Explained* (Boston: Little, Brown, 1991), chs. 5–6. His later shorthand, "fame in the brain," emphasizes that no central presentation makes a content conscious; distributed effects on memory, report, reasoning, and control give some contents a wider career than others. The chapter uses that architecture as a positive resource without treating Dennett's rejection of

additional intrinsic qualia as a denial of pain or experience; ch. 12 and the Appendix preserve content-bearing discriminative states and explicitly affirm the reality of pain. The house identity claim remains a distinct commitment. First-person reports remain evidence about what appears to a subject, handled under the heterophenomenological discipline stated in Chapter 6. Ned Block, "On a Confusion about a Function of Consciousness," *Behavioral and Brain Sciences* 18, no. 2 (1995): 227–247, distinguishes phenomenal consciousness from access-consciousness. The proposed coincidence of phenomenal presence with the distributed access-and-control profile is therefore a defeasible hypothesis, not an equivalence established by definition. Testing compares fallible first-person evidence with separately measured availability and control; the use of reports as evidence does not make report a necessary condition.

**13b.** David J. Chalmers, "Strong and Weak Emergence," in *The Re-Emergence of Emergence*, ed. Philip Clayton and Paul Davies (Oxford: Oxford University Press, 2006), secs. 1–2, distinguishes unexpected but deducible high-level phenomena from strong emergence and treats strongly emergent qualities and strong downward causation as separable. His initial formulation uses in-principle deducibility; note 1 allows finer distinctions between conceptual and metaphysical necessitation. The main text does not adopt non-deducibility alone as an anti-physicalist test, since the house type-B position distinguishes those necessities. Brian P. McLaughlin, "The Rise and Fall of British Emergentism," in *Emergence or Reduction?*, ed. Ansgar Beckermann, Hans Flohr, and Jaegwon Kim (Berlin: de Gruyter, 1992), 49–93, supplies the historical background. Tim Crane, "The Significance of Emergence," in *Physicalism and Its Discontents*, ed. Carl Gillett and Barry Loewer (Cambridge: Cambridge University Press, 2001), argues that looser uses mark no stable boundary from ordinary physicalism.

**14.** The "hostage" framing names a dependency rather than an argument for identity: a phenomenal-concept explanation can show why a gap would persist *if* a psychophysical identity holds, but it cannot establish the identity. Block and Stalnaker's narrower claim is that the failure of *a priori* analysis does not establish an unclosable explanatory gap; see "Conceptual Analysis, Dualism, and the Explanatory Gap" (cited n. 11), esp. secs. 1 and 4. Ned Block's "The Harder Problem of Consciousness," *Journal of Philosophy* 99, no. 8 (2002): 391–425, esp. 391–95 and 419–25, raises a related metaphysical pressure through the other-minds problem for functionalism: how much of the mind–body problem rides on the metaphysics of the identity rather than on the epistemology of our concepts. The book accepts the dependency as a true description of what the deflation rests on, not as an objection to it. This dependency does not privilege historically wide identity over structural identity. Either identity could support the no-additional-product argument, whereas necessary grounding requires a different stopping argument. The evidence ranking in Chapter 6 therefore carries through unchanged: support for an intentionalist account of ordinary contrasts exceeds the support achieved for the specifically historical extension.

**15.** John R. Searle, *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992), esp. chs. 1 and 5, and *Mind: A Brief Introduction* (New York: Oxford University Press, 2004), where "biological naturalism" is stated: consciousness is a higher-level feature of the brain, *caused by* lower-level neurobiological processes and *realized in* the brain as a system, no more mysterious in principle

than liquidity or digestion. Searle's cure for the "Bad Argument" in the philosophy of perception is adopted in Chapter 6; the present chapter's disagreement is confined to the hard problem.

**16.** The structural point is that Searle's "caused by and realized in" is officially non-reductive: he holds consciousness to be a system-level biological feature, denies that it floats causally over and above the neurobiology, and also insists on a first-person ontology that cannot be reduced to third-person facts (*The Rediscovery of the Mind*, ch. 5). He presents biological naturalism as avoiding both dualism and materialism, so the disagreement should not be forced into Searle's classification as a physicalist. It concerns whether causal realization plus first-person irreducibility can perform the work this book assigns to reductive identity. The chapter's answer is no: Searle can keep one biological system and reject dualism, but without strict identity he cannot use identity as the terminal point that retires the bridge demand.

**17.** The content-deferral note, posted openly in Part Two and addressed in Part Three, concerns semantic address rather than every condition hidden in "of the right kind." Chapter 15 develops a teleosemantic model (Millikan, Dretske) of world-directedness and misrepresentation, with worked perceptual and bodily-address cases. Evaluative pain content, mood's diffuse accuracy conditions, and the warrant for experience-directed relief remain incompletely explained. Chapter 6 separately treats phenomenal mode and state consciousness; Chapter 17 treats the persisting subject; Chapter 24 applies the distinctions to candidate machine bearers. See Chapter 6, n. 19, where the aboutness debt is marked from the other side.

**18.** David J. Chalmers, "Phenomenal Concepts and the Explanatory Gap," in *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, ed. Torin Alter and Sven Walter (New York: Oxford University Press, 2007), 167–194, esp. secs. 3–4. Chalmers requires both C and E to receive topic-neutral formulations. If P&~C is conceivable, C lacks transparent physical explanation; if it is not, the zombie has C, while P&~E remains conceivable because corresponding beliefs can differ in truth, justification, or cognitive significance. Section 4, Option 2 explicitly considers narrow explanations of acquisition, inferential disconnection, and belief, then augmented explanations of reference and knowledge. The text accepts the limit rather than claiming that an identity alone answers the dilemma. Helen Yetter-Chappell, "Dissolving Type-B Physicalism," *Philosophical Perspectives* 31 (2017): 469–498, doi:10.1111/phpe.12099, argues that a possible knower with general physical knowledge, phenomenal concepts, and suitable inferential wiring can know psychophysical identities a priori. Her conclusion challenges the ideal-gap thesis of type-B physicalism and supports a type-A phenomenal-concept strategy. The distinction in the reply between experiential concept acquisition and the evidential basis of a later judgment is the book's own: enabling acquaintance does not alone settle whether the judgment is a priori. Nor does an architecture-relative acquisition account establish a universal ideal gap.

**19.** Philip Goff, "A Posteriori Physicalists Get Our Phenomenal Concepts Wrong," *Australasian Journal of Philosophy* 89, no. 2 (2011): 191–209, doi:10.1080/00048401003649617, distinguishes transparent concepts (complete nature disclosed), translucent concepts (part disclosed), mildly opaque concepts (only contingent identifying features disclosed), and radically opaque concepts (no nature or identifying feature disclosed). His challenge needs only translucency: if thinking of pain in terms of how it feels reveals something of pain's nature, a posteriori physicalism must

explain how that disclosure remains compatible with an identity not knowable a priori. The text grants access to manifest world- or body-presenting character, not a grasp of essential nature sufficient for a priori identification under physical concepts. Goff's translucency argument needs only partial revelation, not complete essence-revelation; the reply therefore denies the inference from disclosed character to conceptual identification of its physical nature. If translucency entails that stronger conceptual disclosure, the house account does not accept it. That distinction between presentation and essence-grasp is the book's response, not Goff's concession. The wider dispute turns on what introspection warrants. Ned Block's "Mental Paint," in *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, ed. Martin Hahn and Bjørn Ramberg (Cambridge, MA: MIT Press, 2003), 165–200, Amy Kind, "What's So Transparent About Transparency?", *Philosophical Studies* 115 (2003): 225–244, and Martine Nida-Rümelin, "Grasping Phenomenal Properties," in *Phenomenal Concepts and Phenomenal Knowledge*, ed. Alter and Walter (New York: Oxford University Press, 2007), 307–338, press stronger intrinsic-quality or essence-disclosure readings. The intentionalist reply relies on comparative explanation, not an introspective proof that the rival posit is absent.

**20.** Thomas Nagel, *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature Is Almost Certainly False* (New York: Oxford University Press, 2012), ch. 3, esp. 38–44, 47–52, and 58–60. At 39–41 he presses the missing subjective aspect and the apparent contingency of psychophysical identity; at 41–42 he allows that a necessary, nonconceptual connection may be hidden by our inadequate concepts. At 43 he expressly presents the framework argument as conditional on psychophysical reductionism failing. At 50–52 he separates explaining which organisms have conscious life from explaining their historical emergence; at 59 he allows distinct combinations of constitutive and causal, teleological, or intentional historical accounts. The reply therefore neither attributes an unconditional rejection of every identity to him nor infers an evolutionary history from its own constitutive thesis. Nagel's skepticism is not religiously motivated and commits him to no definite alternative. The 1974 bat paper, "What Is It Like to Be a Bat?", *Philosophical Review* 83 (1974): 435–450, leaves the metaphysical verdict open; see Chapter 9, n. 1.

**21.** Substance dualism (Chapter 9) posits a second kind of substance. Panpsychism and phenomenal Russellian monism (Chapter 10) make phenomenal properties fundamental; protophenomenal Russellian monism instead posits an intrinsic basis that constitutes experience without itself being phenomenal. For that distinction see Chalmers, "Consciousness and Its Place in Nature" (cited n. 12), sec. 11. Russellian theorists may regard this basis as the physical world's underlying nature, not an addition to it. The book's disagreement concerns the need for a fundamental phenomenal or protophenomenal basis beyond the structural-relational account it defends. Calling that basis an "addition" expresses the book's explanatory assessment, not an agreed description of the rival. On the house identity, consciousness is a high-level relational achievement of physical systems without such a further basis. The Coda, *The Marvel of Conscious Existence*, opens with *The Singularity Is Here* — the arrival of machines that merely seem conscious — and closes on what the cleared picture leaves standing: a being with a stake in a world that can cost it everything.

# PART THREE — Meaning, Self, and Other Minds

## Chapter 12: Space, Time, and the Possibility of Reference

> *"Hence, as things are, particular-identification in general rests ultimately on the possibility of locating the particular things we speak of in a single unified spatio-temporal system."*
>
> — P. F. Strawson, *Individuals* (1959), 38

---

CHAPTER OVERVIEW

Imagine two perfectly matched stones and ask how a thought could pick out this one rather than its twin. No list of shared traits can do the job, since the same list fits both, and no pattern inside a mind can give itself an address in the world. I argue that thought about a particular thing needs a link that singles it out, though the link need not pass through human eyes or ears. Meeting the thing in shared space and time gives us the clearest case; names, records, and other people's words can then pass the link to someone who never meets it. A machine too may gain a route to that stone through real contact or a past that reaches beyond its own form, but the route alone does not show that it understands.

> Put two stones on the table, matched in mass, color, and shape. Ask someone to pick up the one you mean. "The smooth gray stone" will not help; both qualify. "The one beside this lamp" will, because now you have supplied a relation to something already picked out. A longer list of shared qualities can leave us very precisely undecided.

The puzzle concerns thought about *that particular thing*, what philosophers call *de re* thought. We usually solve it without noticing: point, look, trace a name back to someone who encountered its bearer. Now remove all those routes. Imagine a mind with no history and no channel to a world, equipped with every formal dictionary entry but no encounter, teacher, or inherited name. What could make its thought concern one stone rather than the other?

A language model does not fit that description. It inherits human texts and a training history. The question for machines concerns what those routes transmit and how the system uses them, not whether a real model amounts to a historyless mind.

Part Two argued that experience consists in representational content of the right embodied, world-directed kind. Part Three now asks how a state comes to be about a world at all. Biological function, causal history, and a community of speakers help fix content. Singular reference adds a specific constraint: a concrete particular needs a relational handle that distinguishes it from qualitative duplicates. Encounter can supply that handle; later consumers can inherit it through the practice.

The constraint belongs to a referential practice, not every speaker's sensory biography. Somewhere in the chain, a world-involving relation must distinguish the particular from its rivals. Strawson and Gareth Evans developed the modern account of how our practice achieves this; the rationalist challenge asks whether thought could instead reach a particular through description alone.[1]

### 12.1 Why Reference Seemed Free

The inner theater makes reference look free. To think about a stone, on this picture, is to entertain an inner item that pictures or encodes it. The stone causes the item; after delivery, the thinking goes on inside. Cut every sense and the item remains, apparently carrying its aboutness with it.[2]

But an inner item, considered merely as a pattern or state, fixes no object by itself. Relations to the world distinguish this stone from its rivals, and a relation is not luggage the mind can carry indoors. The sealed-mind question asks whether singular reference can originate without encounter, inheritance, or another channel that selects one particular. It cannot.

### 12.2 Tokening and Singling Out

Distinguish a system's *having* a representation from a token's *picking out a particular*. A flashcard bearing "the Eiffel Tower" carries a public representation through the practice that produced it; the ink alone does not fix the tower. A model emitting the same string has tokened a well-formed item, richly connected to others, and training may give its internal mechanisms task-shaped structure carrying inherited content. Whether any state singles out that tower in Paris remains a further question.

Call the first achievement *tokening*: producing or entertaining an item with the relevant representational form and internal liaisons. Call the second *singling out*: standing in the relation to a particular that makes a state about *that* one rather than another. Neither internal structure nor qualitative fit can single out *this* stone over its perfect twin; some discriminating world-involving relation must tie a state or practice to one.

Tokening comes cheaply to a sealed mind or a model, though training may earn the model more. Singular reference also needs a world-involving relation.

### 12.3 The Floor in Three Steps

1. **Step one: qualitative fit alone does not ground de re reference to a concrete particular.** The stones become distinguishable when we borrow an anchor: *this* lamp, *this* speaker, *here*. Remove the borrowed anchors, as Max Black's perfectly symmetrical spheres do, and a complete qualitative description hangs over both at once.[3] A definite description may still denote a unique satisfier in an ordinary asymmetric case. The narrower claim is that qualitative, uncentered fit cannot by itself determine which member of a symmetric pair this thought is de re about.

   Centered, indexical, and causal descriptions do not refute that claim; they reveal its metasemantic point. "The stone two feet from me" discriminates because *me* is evaluated from an occupied standpoint. "The cause of this very use" discriminates by tracing this token's actual ancestry. The expressions can remain grammatically descriptive, but their anchor is fixed by

standing in a relation, not by matching a further qualitative specification. Any regress of descriptions ends in a token, center, or producer already placed. Causal contact, or a name inherited from someone who once had such contact, supplies a handle of that kind. Strawson and Evans make spatiotemporal location fundamental to our practice because it supplies the public framework in which those routes converge.[45] The moral is narrower than "every referrer must locate." A singular practice must distinguish.

2. **Step two: a referential practice gets placed through world-involving producers.** A perceiver does not infer its perceptually present *here* and *now* from a neutral description; bodily and sensory encounter places it in a world already under way. That encounter supplies the clearest producer relation: *this stone*, presented here. Only now do the little words earn their reach. *Here*, *left*, and *now* single out from a standpoint: a demonstrative borrows its aim from the encounter that anchors it — a relation the twin does not stand in, not a description the twin might also satisfy. But the relation can travel. A child who learns *Napoleon* from teachers need not meet Napoleon; testimony and deference connect the child's use to producers whose practice was anchored through records, artifacts, and earlier encounters. Thought alone cannot manufacture the initial singular element, but an individual thinker can inherit it.[6] Sensibility therefore names the paradigmatic entrance of particulars into a practice, not a demand that every later consumer personally sense each object.

3. **Step three joins them.** Qualitative description cannot distinguish duplicates (step one). A world-involving producer relation can establish a singular handle, and a causal-historical practice can pass it to consumers (step two). Therefore a closed, historyless system cannot create singular reference from internal form alone. It may possess the form of a spatial representation, a superb atlas with no route connecting any mark to this particular terrain. If it later receives an inherited name, a structured feed from a real environment, or another discriminating relation, the verdict can change. The floor is ecological and genealogical: somewhere in the practice, the world must get a vote. The same lesson will return when instantaneous physical duplicates diverge in content because their histories differ: present internal structure does not exhaust the relations that fix what a state is about.

This floor concerns singular address; general property content and abstract reference face different determination problems.

## 12.4 Objections

### 12.4.1 The mathematician's numbers

We refer constantly to things no sense could ever touch: the number seven, the empty set, the value of justice, the proposition that snow is white. None of these occupies a region of space or a moment of time. If reference required spatiotemporal location, none of it would be possible — and plainly it is. So the thesis is false.

The objection reaches beyond the thesis, which concerns concrete de re reference, where a qualitative twin can stand outside the description and be missed. Abstract reference raises a different determination problem. In our practice it grows from symbol use, proof, teaching, stipulation, and applications embedded in a public world: seven from traffic with pluralities, the empty set

from operations on collections, a proposition from worldly conditions that would make it true. The practice remains world-embedded even when its referent lies outside space and time; structural, inferential, or stipulative relations may fix an abstract referent without causal contact with it.[7] That genealogy leaves room for nominalists, structuralists, and non-haecceitistic Platonists. A Platonism demanding singular reach to *this* causally isolated abstractum among structurally equivalent candidates inherits a further problem. Even seven was raised on apples.

### 12.4.2 The self-sufficient intellect

The rationalist asks why the limits of human cognition should bind every mind. Perhaps an intellect could grasp its objects directly without sensory contact: the tradition's *intuitus intellectualis*, or intellectual intuition. Would our account rule it out merely because it lacks familiar senses?

The objection has a real historical champion. Leibniz held that every individual thing has a *complete individual concept*: a description that includes everything ever true of it.[8] Such a concept, he thought, fastens on Caesar and no other — not by locating him, but by the sheer completeness of the description. Leibniz, a mathematician of the first rank who thought the world through to its logical floor, gives the rationalist his strongest card.

A description, however complete, may determine a unique satisfier; Evans's distinction asks whether the thinker thereby achieves the object-dependent reach at issue here.[9] Leibniz can deny that a qualitative duplicate remains possible and insist that Caesar alone satisfies Caesar's complete concept in every possible world. Even then, we can ask what attaches this token of thought to Caesar rather than merely making its description fit him.

Leibniz's best modern reader would object that this sells the doctrine short. Robert Merrihew Adams, whose 1994 study set the standard for taking Leibniz's metaphysics whole, reads the complete concept as no mere list of qualities: each substance expresses the entire world from its own point of view, so the concept arrives already relational, already perspectival. Grant the reading; it moves the problem without solving it. A description of a perspective tells you what the view contains, not what attaches this token of thought to its subject. If a primitive thisness, divine selection, or direct presentation supplies that attachment, the singular element has been supplied rather than derived from qualitative content. The rationalist may call the relation direct intellectual presentation. Fine; then the view has supplied a discriminating presentation rather than generated reference from description alone. The concession keeps its teeth: a faculty that reaches its object owes the same account any channel owes, of how it discriminates this one, while a faculty that supplies no such relation remains internal form. What matters is the reach, not whether it arrives through a human sense organ.[10]

### 12.4.3 Testimony, Disability, and Unfamiliar Channels

The personal-sensory version of the thesis would prove too much. A blind or deaf thinker refers fully; a speaker can refer to Napoleon by testimony; a scientist can name an entity detected only through instruments. The ecological version predicts all three. Reference can pass through touch, records, teachers, cables, and instruments because the practice contains producers and preservative chains. A brain receiving structured signals from a real environment may acquire singular handles through the cable; a machine may acquire them through unfamiliar channels.[11]

The condition fixes on a discriminating world relation somewhere in the practice and stays neutral on hardware. What it rules out is not silicon or disability. It is self-anchoring form.

### 12.4.4 The machine that has space and time of its own

A modern language model, the fourth objection says, *has* space and time of its own. It processes inputs in sequence, token after token, giving it a kind of time. It also organizes meanings in a high-dimensional vector space where nearness encodes similarity, giving it a kind of space. The model therefore commands a spatiotemporal framework, and the floor poses no barrier.

The model's "space" and "time" do real work, but not all the work reference needs. A token sequence orders symbols; an embedding space maps similarities and, through training, acquires task-shaped structure and learned sensitivity to textual distributions. Neither geometry alone distinguishes this stone from its twin, because qualitative and structural relations can survive a swap of distal particulars. Internal order can make inherited content causally active without fixing a singular address or originating the semantic standard. The inference from interpretable geometry to self-baptized worldly relata therefore fails. Training, retrieval, feedback, tools, or sensors offer stronger candidate routes, and some may meet the floor by linking a state to a particular.

The strongest version says the geometry may model the *world*, not merely words. A game-playing model can encode the current board state, and interventions on that representation can change its predictions. Grant the finding: a learned world-model can carry content fixed by training and use. A formal game position needs no geographical address to count as content. If we ask which particular physical board or episode the state concerns, however, we ask a further question. That singular link must come from training history, a live feed, inherited producer chains, or some other route to the particular.

Discourse can preserve a producer chain for a consumer; a live retrieval, tool, or sensor call can transduce a particular's state at use. The floor classifies these possibilities rather than barring them. Even when an episode fixes a referent, two questions remain: whether the system participates as a consumer rather than routing a referential token, and whether it integrates the content across memory, correction, inference, planning, learning, and action. Chapters 15 and 23 meet them. Coordinates can carry content, but geometry alone does not baptize its relata.[12]

## 12.5 A Real Space and Time, and the Attribution Question Left Open

Concrete de re reference needs a relation that distinguishes one particular from its qualitative rivals. Encounter supplies the clearest producer route; history and community can carry it to consumers. The requirement concerns genealogy, not hardware or private sensory biography.

The argument does not make space and time *mind-made*. Sensibility supplies a mode of access to a spatiotemporal world, never the world itself. Stones stand at a distance whether or not anyone perceives them; the senses supply a standpoint *in* that real frame. Strawson reconstructed this realist reading from the older argument, and the externalism ahead requires it.[13] A book that insists you see the apple rather than an inner picture cannot turn space into another inner picture.

Unlike the sealed mind, a language model inherits causal history, public meanings, and task-shaped internal structure. Its tokens may carry derived routes to particulars, and a deployed system may participate in some. The remaining questions ask whether it acts as a consumer by deferring to producers, using the token as part of the practice, and remaining corrigible by its standard, and whether it integrates the content across memory, correction, inference, planning, learning, and action. Part Four takes up those questions; the model is not simply the sealed mind with better arithmetic.[14]

**Chapter Summary**

**The claim.** Concrete reference to this person, tree, or place cannot arise from qualitative fit or internal structure alone. It needs a route through the world that singles out one particular; encounter provides the clearest case, while testimony and causal history can pass the handle onward.

**Where the argument stands.** An uncentered description can fit perfect duplicates, so added internal detail cannot select one of them. Centered and causal descriptions work only by relying on a standpoint, token, or source already fixed through a relation to the world. Spatiotemporal contact can provide that anchor, and public practice can preserve it for later users; the requirement concerns history and organization, not hardware, so silicon can refer and each human speaker need not sense the object firsthand.

**What's left open.** Abstract reference may need another account. A corpus history may link an artificial state to a distant object or only to linguistic contexts, and even successful reference leaves the system's understanding unsettled.

**The hand-off.** A signal can point home without knowing where home lies. The next chapter separates information that travels from meaning that reaches.

**Footnotes**

**1.** The argument's modern, realist statement is Strawson's and Gareth Evans's (notes 4 and 5): material bodies in a single spatiotemporal system are the basic particulars our referential practice rests on, and a fundamental producer route singles one out by locating it in that system. Evans's producer-consumer distinction also explains how later speakers inherit singular reference without encountering the object. The distant ancestor is Kant's claim, in the Transcendental Aesthetic of the *Critique of Pure Reason* (1781/1787), that space and time are the forms under which a sensible mind is given particulars. This chapter keeps the realist and practice-level insight while leaving behind both Kant's idealism and the stronger claim that every individual act of reference requires a personal sensory presentation.

**2.** The genealogy is the work of Chapter 2. The semantic version of the inner-theater error — locating aboutness in an inner item rather than in a state's relation to the world — is the same reification the book diagnoses for experience in Chapters 5 and 7, transposed from phenomenal character to intentional content.

**3.** Max Black, "The Identity of Indiscernibles," *Mind* 61 (1952): 153–164, argues against the Identity of Indiscernibles through a symmetrical two-sphere universe. The chapter makes no claim that

spatial position metaphysically grounds their distinctness; in Black's world the positions are themselves descriptively symmetrical. Its cognitive prediction is that an equally symmetrical thinker, with no asymmetric relation to either sphere, could not think de re of one rather than the other. Demonstrative location is the paradigm of the needed handle, but a causal-historical name or inherited producer chain can preserve the same distinction.

**4.** P. F. Strawson, *Individuals: An Essay in Descriptive Metaphysics* (London: Methuen, 1959), Part I. Strawson argues that material bodies in a single spatiotemporal system are the *basic particulars* of our conceptual scheme. The claim concerns the structure of an identifying practice, not a requirement that each user personally perceive each particular. "Descriptive metaphysics" is his term for laying bare the structure ordinary thought already has, as against replacing it with a revisionary scheme.

**5.** Gareth Evans, *The Varieties of Reference*, ed. John McDowell (Oxford: Clarendon Press, 1982), esp. ch. 4 for the *fundamental Idea*, ch. 6, "Demonstrative Identification," and ch. 11 on name-using practices. Evans's fundamental Idea locates an object in objective space and time; his distinction between producers and consumers explains how a name-user can retain singular reference by deferring to a practice whose producers possess direct information links to the object. The first relation anchors; the second preserves and distributes the reach.

**6.** Evans distinguishes purely descriptive thoughts from information-based identification and object-dependent thought (*Varieties of Reference*, chs. 1 and 4–6). He also recognizes descriptive names and non-Russellian singular thoughts; not every thought expressed with a singular term has the object-dependent character at issue here. The chapter draws on his account of that narrower achievement, rather than treating the singular/general division as exhaustive. Perceptual and bodily anchoring supplies the fundamental producer case; Evans's later producer-consumer apparatus allows individual thinkers to inherit that reach. The older form is Kant's contrast between concept and intuition (*Critique of Pure Reason*, A51/B75, in the Meiklejohn translation), with the proviso of note 1: the chapter keeps the need for a singular presentation somewhere and drops both the idealism and the personal-sensory universal.

**7.** Strawson argues in *Individuals*, ch. 1, that reference to non-basic particulars is introduced within a conceptual scheme whose basic identifying practice concerns spatiotemporal bodies. The extension to abstracta is the chapter's genealogical proposal, not a result Strawson alone establishes. It says how a public practice acquires and teaches its mathematical symbols, not that the number seven occupies space or that every mathematician must trace each proof back to apples. Paul Benacerraf, "What Numbers Could Not Be," *The Philosophical Review* 74, no. 1 (1965): 47–73, especially sec. II, develops the familiar identification problem: structurally equivalent reductions of numbers leave a haecceitistic Platonist owing an account of why the practice reaches one abstract object rather than its rival.

**8.** The doctrine of the *complete individual concept* (*notio completa*) is Leibniz's: every individual substance has a concept so complete that it contains, and allows the deduction of, every predicate ever true of that substance — the position stated in the *Discourse on Metaphysics* (1686), sec. 8 (no verbatim quotation is made here; for the text see G. W. Leibniz, *Philosophical Essays*, ed. and

trans. Roger Ariew and Daniel Garber [Indianapolis: Hackett, 1989]). The complete concept is the strongest historical attempt to individuate a particular by descriptive content alone, with no spatiotemporal handle, which is why it is the rationalist's best instance of the *intuitus*-style grasp the objection invokes. Robert M. Adams, *Leibniz: Determinist, Theist, Idealist* (New York: Oxford University Press, 1994), chs. 1–3, treats contingency, identity, and complete concepts; Part III, chs. 9–13, treats the wider monadic and perspectival system. Adams supplies the strongest reconstruction rather than an objection to the chapter. The revised reply does not assume Leibniz must admit duplicates: even necessary unique satisfaction leaves open what attaches this token of thought to the satisfier.

**9.** The reply draws on Evans's account of object-dependent thought (*Varieties of Reference*, chs. 1 and 4–5; see note 6) in assessing the complete concept. A fully specified general concept may denote a unique satisfier while failing to place the thinker in an object-dependent relation to it. What closes that gap is a singular element: demonstrative presentation, causal acquaintance, baptism, or an inherited chain preserving one of those relations. The structural objection runs back through Russell's distinction between knowledge by acquaintance and by description, "Knowledge by Acquaintance and Knowledge by Description," *Proceedings of the Aristotelian Society* 11 (1910–11): 108–128. Causal descriptivism can build a relation into a description, as in a token-reflexive clause such as "the object at the causal origin of this very use": David Lewis, "Putnam's Paradox," *Australasian Journal of Philosophy* 62 (1984): 221–236, who coins the label; Frederick W. Kroon, "Causal Descriptivism," *Australasian Journal of Philosophy* 65 (1987): 1–17. Lewis distinguishes a causal description from a genuine causal constraint and supplements his preferred causal descriptivism with the eligibility, or naturalness, of referents (225–228). The claim here concerns token-reflexive versions whose success depends on the actual use and its actual causal ancestry: grammatical descriptiveness does not make such a route purely qualitative and uncentered. Those versions can preserve descriptivist semantics while meeting a relational constraint. This limited compatibility does not show that every causal-descriptivist theory accepts the chapter's ecological metasemantics.

**10.** The *intuitus intellectualis* — an intellect that would create or grasp its objects directly, without being sensibly affected — is Kant's foil for finite cognition (e.g., B71–72, B135, B145). Kant denies humans possess it. The reply here is conditional rather than theological: if direct intellectual grasp supplies a genuine relation to one object rather than its duplicate, it meets the chapter's world-involving constraint by another route. What it cannot do is derive singularity from purely general descriptive content. Nor does the conditional trivialize the constraint: either the posited grasp stands in a discriminating relation to its object, in which case it functions as a channel and its defender owes an account of how the channel discriminates, or it operates on internal content alone, in which case it hangs over duplicates with every other description. "World-involving" names the relation's achievement, not a courtesy title awarded to whatever succeeds. Cf. Evans, *Varieties of Reference*, ch. 4.

**11.** The organ- and substrate-neutrality of the anchoring relation mirrors the book's standing anti-chauvinism. Chapter 23, meeting Mollo and Millière, grants that sensorimotor transduction is not necessary for reference; causal-historical learning, social deference, and other channels

may do the work. The present chapter adds the ecological constraint: a singular practice cannot originate its distal addresses from internal description alone. Some producer or learning route must be affected by, discriminate, or otherwise become accountable to the particular; consumers may inherit that relation. The brain-in-a-vat and machine cases are therefore decided by the feed and history, not by the familiar shape of an organ.

**12.** The "merry-go-round" is Harnad's symbol-grounding image (Chapter 23). An internal metric or sequence can support genuine computation and, when shaped by training and consumer use, task-specific causal organization and content. What geometry alone cannot do is fix a distal singular referent: structural relations remain invariant under permutations of the worldly relata. Training history, a live indexical feed, or an inherited producer chain may add the missing relation. The empirical case is Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg, "Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task," *International Conference on Learning Representations* (2023), arXiv:2210.13382: board state is recoverable from Othello-GPT's activations, and intervention on the recovered representation changes subsequent predictions. The live philosophical dispute divides accordingly. Steven T. Piantadosi and Felix Hill, "Meaning Without Reference in Large Language Models" (2022), arXiv:2208.02957, defend conceptual-role meaning without reference; Matthew Mandelkern and Tal Linzen, "Do Language Models' Words Refer?", *Computational Linguistics* (2024), arXiv:2308.05576, argue that model words may inherit reference through human practice. This chapter does not settle whole-system understanding; it rules out internal structural readability as self-sufficient metasemantics and classifies candidate inherited routes.

**13.** P. F. Strawson, *The Bounds of Sense: An Essay on Kant's "Critique of Pure Reason"* (London: Methuen, 1966), which separates the defensible Kantian argument (the conditions of a possible experience of an objective world) from the transcendental idealism Strawson rejects. The book takes the former and leaves the latter, consistent with its realism: space and time are real features of the world; sensibility is our mode of access to them, not their author. Evans's externalism (*Varieties of Reference*) likewise presupposes a mind-independent spatial world that thought reaches rather than constitutes.

**14.** The comparison now turns on genealogy rather than sensory inventory. A genuinely historyless sealed mind lacks any route that fixes a singular address. A trained language model does have a causal history: public word-types carry derived meaning, and training gives internal mechanisms learned sensitivities to textual distributions. Evans's consumer condition (note 5) requires more than causal descent: the use defers to producers, manifests participation in the relevant practice, and remains corrigible by its standard. These achievements come apart: a token's inherited derived reference, a mechanism's causal internalization of that content, an agent's consumer participation and integrated understanding, and an autonomous individual's stake, vulnerability, and welfare. Chapters 14–16 distinguish these achievements; Chapters 23–24 apply them to artificial systems.

## Chapter 13: The Difference Between Information and Meaning

> *"These semantic aspects of communication are irrelevant to the engineering problem."*
>
> — Claude Shannon, "A Mathematical Theory of Communication"


CHAPTER OVERVIEW

A thermometer can help us learn the room's temperature without knowing anything about cold, and its reading can mean something even when no one looks. I want to separate three things here: the way the signal tracks heat, what makes it a reading of that heat, and a person's grasp of what to do about it. A device can fail to give a reading, give a wrong one, or give one that a doctor reads correctly but fails to understand. More bits alone cannot explain those different failures, since the standard for getting something right must come from somewhere. I argue that content needs such a standard, whether for a claim that can be false or an order that can go unfulfilled, and ask how history and use can fix it.


A thermostat hangs on the wall of nearly every room I have worked in. It registers temperature, compares the signal with a set point, and switches the heat. In a loose sense, it "knows" and "represents" the room's condition.

Does the thermostat understand what cold means?

Its reading can carry derived content: calibration and use make it a reading of the room's temperature. That grants the device more than a door responding to a push, but not an understanding of cold. Systems that write essays, pass professional exams, and describe cities they have never stood in raise the question in harder form.

### 13.1 The Gap Between Carrying Information and Meaning It

Three achievements need separating: correlation, content, and understanding. In the engineering sense Claude Shannon made precise, information measures uncertainty and its reduction across a channel. It does not measure how much anyone understands. A thermometer's changing state can reduce uncertainty about temperature without explaining what makes that state a temperature reading.

Now compare the thermometer with a doctor reading 102.6 degrees Fahrenheit. Physical sensitivity supplies the correlation. Calibration, a scale, and a measuring practice give the display content: this temperature, in these units. The doctor can also connect the reading with symptoms, question its reliability, repeat the measurement, and change the treatment. Understanding appears in those capacities working together, not in the addition of another decimal place.

Error sharpens the distinctions. A blank display may signal a malfunction without displaying a false temperature. A working display can misrepresent the temperature when the calibration slips. A doctor can read the number correctly yet misunderstand its significance. The device's derived

content already permits the second failure; understanding need not wait inside the thermometer for error to become possible.[1]

Where do the standards come from? A designed scale borrows them from a practice. Chapters 14 and 15 ask how a producer-consumer history can also fix standards independently of an interpreting practice. Maps can misrepresent a route; instructions can go unfulfilled; beliefs can be false. These different assessments concern content, while the question of who understands by means of it remains further. Content alone establishes neither a subject nor consciousness, free agency, autonomy, or welfare.

### 13.2 The Symbol Grounding Problem

Stevan Harnad named the symbol grounding problem in 1990 and gave it an image. Here is an adaptation of his harder dictionary case.

Imagine trying to acquire a first language from a Chinese-Chinese dictionary alone, with no already-grounded concepts to help. You find an entry, but its definition consists entirely of other Chinese characters. You look those up. More characters. You could keep circulating through the dictionary without establishing what any character means. Someone learning Chinese as a second language brings concepts and worldly experience to the task. Harnad's harder case removes that resource: symbols defined solely through other ungrounded symbols supply no way off the circle.[2]

More symbols lengthen the dictionary; sharper rules speed formal operations. Neither, by itself, supplies the missing connection. Dretske's cut is the one I want here. Thermometers, maps, and stories mean by our purposes; change the practice and you change or erase what they say. That is derived meaning. Original meaning is not that assignment, and not an inner spark. It is a state's function of indicating — the place, as he put it, where the misrepresentational buck stops. Once we have that capacity we can adopt Dennett's intentional stance; the stance expresses intentionality, it does not analyze it. Chapter 14 develops the distinction without treating either kind of content as counterfeit.[3]

Harnad grounded elementary symbols in sensorimotor traffic so higher ones could inherit a worldly connection. That need not supply the only route, and reference does not yet amount to understanding. The narrower lesson holds: a worldly address requires a relation to what the symbols pick out that is not another symbol.[4]

First, *information* must name one thing at a time.

### 13.3 Information in Shannon's Sense

A single word has carried two loads. Shannon's 1948 paper measures information in bits, allowing engineers to study transmission without first settling meaning. It also draws that boundary almost at the outset: the *semantic* aspects of communication, he wrote, are irrelevant to the engineering problem.[5] Whether the message means anything, whether it is true, whether it concerns the weather or the war: none of it enters the mathematics. Compare equal-length strings over the same alphabet. A source choosing each character independently, with all characters equally likely, has greater expected information than an English source whose patterns make later characters predictable. This compares average uncertainty, not the information of any two particular pages.

The measure depends on the ensemble the text was drawn from; change that ensemble and you change the measure. Neither number tells us what the text means.

Dretske's 1981 information is already something else. A signal carries the information that *s* is *F* when, given the signal, that condition is nomically settled. Analog and digital encodings mark how that specific content is nested, not how many average bits an ensemble contains. Shannon's quantity is the engineering floor. It is not what Dretske then tried to turn into meaning.

Could meaning be *built* from that richer material, directedness assembled out of information given enough further structure? Fred Dretske made the most serious attempt. In *Knowledge and the Flow of Information* (1981), he took Shannon as the floor and tried to raise genuine content on his own informational relation: a state carries information about whatever it lawfully indicates, and a *learning history* then selects, from those many correlates, the one condition the state comes to mean.[6] Dretske never held that covariation by itself suffices. His thesis joins information to function, and the learning period does the picking. During it, a state gets recruited to indicate one condition among the many it tracks, and from then on that one is what it says. A bare-correlation objection merely repeats his starting point.

A more serious charge runs deeper. Even disciplined information *underdetermines* aboutness, and the recruitment may not fix which condition got recruited. Consider an idealized case in which a fever, its infectious cause, inflammation, a particular thermometer reading, and a worried frown occur together without exception under fixed background conditions. A signal that settles whether the fever occurs then settles each of the other conditions too. Ordinary clinical sequences need not meet that stipulation. Nothing *in the shared informational relation* picks which link the state is about. Recruitment is not more causation along that same line. The state is taken up because it indicated one condition and for a consumer's work — another mechanism that depends on its accuracy. That can make error possible without settling grain: fly or speck, fever or infection.

But the objection rebounds against the proposed solution. Selection history is a causal-historical affair too. A frog's mechanism was shaped in a world where flies and small dark moving specks went together perfectly, and philosophers have spent four decades disputing whether it therefore says *fly* or *small dark moving speck*. If underdetermination sinks covariation, it may sink selection with it.

Two questions have to be kept apart. One asks whether a state can have a correctness condition *at all*; the other asks *which* condition it has when several fit the record. Bare covariation fails the first: a state that merely correlates cannot be wrong about what it tracks. Selection can remove that barrier at the consumer's end, where another mechanism depends on the state's accuracy for its own work. That makes error possible without yet deciding which link in the causal chain the state addresses. Chapter 15 develops the consumer loop; whole-system belief or understanding will depend further on integration, correction, memory, planning, and action.

Shannon gave us a perfect theory of the channel; he did not give us, and never claimed to give us, an account of what travels through it *meaning* anything.

### 13.4 The Same Gap in Brains and in Machines

The same gap opens under two ordinary claims, one about brains and one about machines.

Take the brain first. Open almost any book about it and you will read that the brain *processes information*: it draws information in through the senses, works on it, and ships it out as behavior. The sentence goes down as smoothly as *the kidneys filter blood*. But photons stop at the retina, and pressure waves die in the cochlea. What travels inward is ion flux across membranes. The neurons carry no little envelopes stamped *information*. They fire. Spike trains do carry Shannon information about what the senses met, and neuroscientists measure it in bits per spike. No such measurement supplies what the firing is *about*. Between *the neurons fire* and *the brain processes information about the world*, a semantic description has been added.

Searle's free-mapping objection lands against an unconstrained reading: if any sufficiently complex physical process counts as any computation merely because an observer can map states onto it, computation explains nothing.[7] It does not follow that every computational fact is projected. A mechanism whose causal organization physically realizes the counterfactual structure of a computation can do so observer-independently; Chapter 22 grants exactly that. Yet the semantic question survives the concession. A process can be a genuine computation while what its states are *about* remains unfixed, just as a channel can carry Shannon information without meaning any of it. Calling the brain an information processor may name real causal organization, but it does not by itself ground content. Computational neuroscience keeps everything it has earned.

The machine case poses the same question. We say a language model *represents the world*. A model trained to predict the next token learns each token through the company it keeps, a phrase the linguist J. R. Firth coined in 1957.[8] Firth's slogan became a watchword for *distributional semantics*, which Chapter 23 takes up in detail. Training gives such mechanisms learned sensitivities to linguistic contexts and causally potent internal geometry. Selection and downstream use can make inherited linguistic content indispensable to what those states do. Whether those relations also earn semantic independence requires the closer examination in Chapter 14. Neither causal usefulness nor linguistic ancestry alone settles every further attribution.

The strongest reply presses there. Probe one of these models and structured maps appear in its internal activity: a playable board recovered from a system trained only to predict moves, coordinates of space and time recovered from one trained only on text. More probative findings are interventional. Edit the recovered board and the play that follows changes to match the edit. This is no picture sitting inertly in the weights, but a structure the mechanism *uses*.[9]

That intervention supports a semantic reading as well as showing causal usefulness: the state behaves as a representation of a game position. Representing that formal position needs no geographical *you are here*. Reference to a particular physical board would require a further relation, and understanding the game would require more than one causally useful representation. Chapter 23 examines what the training history and later use establish. The evidence earns attention without deciding every question at once.

The thermostat can keep the room comfortable. It has offered no opinion on winter.

**Chapter Summary**

**The claim.** Information comes cheaply; meaning must earn a worldly address. Covariation, formal relations, and Shannon's measure show structure. None alone makes a state about anything, since none fixes how it could be wrong.

**Where the argument stands.** A thermometer and a doctor may register the same temperature without sharing understanding, and even computation built into a mechanism can provide real causal structure without settling meaning. When a consumer relies on a state to succeed, producer-consumer organization makes error possible. That fact still does not fix what the state represents or at what level; original content needs a world-linked history and use that fix correctness beyond what an interpreting practice supplies, and artificial systems can meet that test in principle.

**What's left open.** Consumer use may narrow whether a frog's state concerns a fly or a small dark moving speck without forcing one answer. Corpus history may carry public reference beyond direct sensorimotor contact while leaving whole-system understanding undecided.

**The hand-off.** Causal structure can carry meaning without creating it. The next chapter asks when the address remains borrowed and when a system earns one of its own.

**Footnotes**

**1.** The possibility of misrepresentation as a constraint on any naturalistic theory of content is a central theme of Fred Dretske's work, from *Knowledge and the Flow of Information* (Cambridge, MA: MIT Press, 1981) to the sharper treatment in "Misrepresentation," in *Belief: Form, Content, and Function*, ed. Radu J. Bogdan (Oxford: Clarendon Press, 1986), 17–36: a state that merely covaries with a condition cannot, by that fact alone, *misrepresent* it, since under a pure covariation reading the state simply indicates whatever it correlates with and can never get that thing wrong. Descriptive representation requires a state with accuracy conditions it can fail to satisfy; directive content instead admits fulfillment or nonfulfillment. Bare covariation alone supplies neither kind of semantic standard. The difficulty of recovering error from covariation (the "disjunction problem") drives the appeal to function and selection history taken up in Chapter 15; for the sharpest statement of how reliable misrepresentation can nonetheless occur, see Angela Mendelovici, *The Phenomenal Basis of Intentionality* (New York: Oxford University Press, 2018), ch. 4, and "Reliable Misrepresentation and Tracking Theories of Mental Representation," *Philosophical Studies* 165 (2013): 421–443.

**2.** Stevan Harnad, "The Symbol Grounding Problem," *Physica D: Nonlinear Phenomena* 42 (1990): 335–346. Harnad's formulation draws explicitly on Searle's Chinese Room as a demonstration that formal symbol manipulation is insufficient for semantics, and extends the point to connectionist systems as well as classical symbol processors. His proposed remedy — grounding elementary symbols bottom-up in sensorimotor interaction with the world (iconic and categorical representations), so that symbols built from them inherit a connection to what they pick out — anticipates the embodied, world-involving account this book develops in Chapters 15 and 17. Harnad is careful to cast his proposal as an account of symbol grounding rather than a full theory of meaning: grounding is necessary for content without being obviously sufficient for it. The strongest recent

challenge to applying his verdict to contemporary systems is Dimitri Coelho Mollo and Raphaël Millière, "The Vector Grounding Problem," *Philosophy and the Mind Sciences* 7, no. 1 (2026): 1–28 (preprint arXiv:2304.01481, 2023), who distinguish five notions of grounding and argue that large language models may achieve the referential variety. Stated at full strength, their case holds that preference fine-tuning can establish world-involving functions, that pretraining alone may ground internal states within limited domains, and — ruling out a simple sensorimotor requirement — that multimodality and embodiment are *neither necessary nor sufficient* for referential grounding. The disagreement concerns what particular training histories and uses establish, not a universal requirement for embodiment. Chapters 23 and 24 assess mechanism content, reference, and whole-system understanding separately; a world-coupled robot supplies one case, not the only possible route to content.

**3.** The conventional/original cut in the body is Dretske's. Instruments, gauges, stories, and pictures represent because our purposes underwrite them; change the practice and you change or erase their meaning. If experiences are representational, they are not conventional in that way: original representations whose power is not derived from agents who already possess it. See Fred Dretske, "Experience as Representation," *Philosophical Issues* 13 (2003): 67–82, at 67–68. In "Misrepresentation," in *Belief: Form, Content, and Function*, ed. Radu J. Bogdan (Oxford: Clarendon Press, 1986), 17–36, at 17–18, he adds that once we have intentionality we can adopt Dennett's intentional stance, but the stance is an expression of intentionality, not an analysis of it; what we are after is "nature's way of making a mistake, the place where the misrepresentational buck stops." Searle's intrinsic/derived taxonomy — *Intentionality: An Essay in the Philosophy of Mind* (Cambridge: Cambridge University Press, 1983), 6, and *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992), ch. 4 — supplies vocabulary the book also uses; it is not the argument of this paragraph. Harnad's "symbol grounding" picks out the same divide from the engineering side. The question of how a physical state could possess original content is the explicit business of Chapter 14.

**4.** Franz Brentano, *Psychology from an Empirical Standpoint* (1874), trans. Antos C. Rancurello, D. B. Terrell, and Linda L. McAlister (London: Routledge, 1995), esp. Book II, ch. 1, where intentionality ("the intentional inexistence of an object") is identified as the mark distinguishing mental from physical phenomena — the thesis introduced at length in Chapter 1. Brentano's formulation seeded both the phenomenological tradition (through Husserl) and the analytic intentionalism this book inhabits (Crane, Tye, Harman, Dretske); for the development and the subsequent misreadings, see Tim Crane, "Brentano on Intentionality," in *The Routledge Handbook of Franz Brentano and the Brentano School*, ed. Uriah Kriegel (London: Routledge, 2017), 41–48. Brentano supplies the directedness to be explained, not a proof that computation cannot realize it. The chapter's further argument concerns explanatory sufficiency: a formal description alone does not settle a state's content, even when a physical mechanism genuinely computes.

**5.** Claude E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal* 27 (1948): 379–423, 623–656. Shannon measures information as the reduction of statistical uncertainty across a channel; in the paper's opening he is explicit that the "semantic aspects of communication are irrelevant to the engineering problem." The technical sense of *information*

and the ordinary sense of *meaning* are two different notions wearing one word — the conflation this chapter is built to undo, and one that recurs elsewhere only by analogy: as the intelligence/understanding distinction in Chapter 24, and as one of the three cuts in Chapter 23.

**6.** Fred Dretske, *Knowledge and the Flow of Information* (Cambridge, MA: MIT Press, 1981), raises content on his own informational relation — a signal carries the information that *s* is *F* when that condition is nomically settled, analog or digital — taking Shannon as the engineering floor, not as the meaning. Technically: the conditional probability of *s*'s being *F*, given the signal, is 1 (while less than 1 on the prior probabilities alone), and Dretske adds a learning period in which a state acquires its meaning by being recruited to indicate one of the many conditions it covaries with. The view is refined in "Misrepresentation" (cited n. 1) and *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995). Burge singles it out as the most rigorous development of "information-theoretic notions alone" (*Origins of Objectivity* [Oxford: Oxford University Press, 2010]). Dretske's information notion derives from Shannon's without being identical to it: he adds the requirement that the conditional probability be 1, which Shannon's average-case measure does not demand, and he distinguishes analog from digital encoding to say when a signal carries a piece of information *as such*. The underdetermination objection pressed here — that lawful covariation runs along an entire causal chain and so cannot, by itself, fix which link a state is about — is the standard pressure point that motivates the teleosemantic supplement (function, selection history) developed in Chapter 15, where Dretske's project is taken up at the place where it succeeds and the place where it needs more than information can give. The objection's reach is the subject of a long dispute the main text acknowledges rather than settles. Fodor pressed it against teleosemantics itself: the frog's fly-detector was selected in an environment where flies and small dark moving specks coincided, and selection, not being an intensional context, seems unable to prefer one description to the other — so the indeterminacy migrates rather than dissolving (Jerry Fodor, *A Theory of Content and Other Essays*, Cambridge, MA: MIT Press, 1990, chs. 3–4). Ruth Millikan reads the content as *frog food there*, fixed by what the consuming systems need in order to succeed; Karen Neander argues for the more proximal *small dark moving thing*, fixed by the response mechanism's own selected sensitivity (Millikan, *Language, Thought, and Other Biological Categories*, Cambridge, MA: MIT Press, 1984, and *White Queen Psychology and Other Essays for Alice*, MIT Press, 1993; Neander, *A Mark of the Mental: In Defense of Informational Teleosemantics*, MIT Press, 2017). The distinction drawn in the main text between the *existence* of a correctness condition and its *determinacy* concedes the residual indeterminacy to Fodor while denying that it levels the two theories: a state its consumers must be able to depend on can be depended on wrongly, which bare covariation cannot manage at all, even where which mistake it made remains open. Chapter 15 returns to the point with the consumer loop in hand — task function, tracking, and the explanatory grain of consumer success — and concedes an irreducible residue wherever rival contents stay explanatorily tied.

**7.** John Searle develops the observer-relativity argument in "Is the Brain a Digital Computer?", *Proceedings and Addresses of the American Philosophical Association* 64 (1990): 21–37, and *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992), ch. 9. His argument exposes the triviality of unconstrained implementation maps: if an interpreter may assign arbitrary state correspondences, almost any physical system implements almost any computation. Chapter 22

distinguishes that free-mapping claim from the stronger conclusion that computation itself is never intrinsic. On the disciplined account adopted there, a mechanism genuinely implements a computation when its physical causal organization realizes the relevant counterfactual structure; that implementation is observer-independent. The present chapter needs only the remaining semantic point: even a real computation does not acquire aboutness merely by computing. Computation fixes a causal organization; further facts about selection, use, and world-involving history must fix what, if anything, its states represent.

**8.** J. R. Firth, "A Synopsis of Linguistic Theory, 1930–1955," in *Studies in Linguistic Analysis* (1957), 1–32, at 11. The essay is the source of the dictum "You shall know a word by the company it keeps." Firth's claim was collocational, a point about the habitual company a word keeps within his broader contextual theory of meaning, and it stands parallel to rather than derived from the *distributional hypothesis* the linguist Zellig Harris had formulated three years earlier: that a word's meaning is a function of its distribution across linguistic environments (Harris, "Distributional Structure," *Word* 10 [1954]: 146–162). Modern distributional semantics draws on both and generalizes them, taking Firth's slogan as its banner; folding his phrase into Harris's program collapses two parallel lineages into one.

**9.** The two probing results are Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg, "Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task," *International Conference on Learning Representations* (2023), esp. secs. 3–4, arXiv:2210.13382 — a transformer trained only on legal Othello move sequences, from whose activations a probe recovers the current board state; and Wes Gurnee and Max Tegmark, "Language Models Represent Space and Time," *International Conference on Learning Representations* (2024), arXiv:2310.02207 — linear probes recovering coordinates of place and date from a text-trained model's activations. The interventional result matters more than the correlational one for the argument in the main text: Li et al. show that editing the recovered board representation changes the model's subsequent legal-move predictions accordingly, which distinguishes a structure the mechanism uses from one an experimenter merely reads off. Nothing in either result addresses which worldly particulars, if any, the state refers to, and the authors do not claim otherwise; Chapter 23 asks what the causal-functional history fixes and whether the containing architecture understands by means of it.

## Chapter 14: Original Intentionality

> *"The study of the mind starts with such facts as that humans have beliefs, while thermostats, telephones, and adding machines don't."*
>
> — John Searle, "Minds, Brains, and Programs"

---

CHAPTER OVERVIEW

A road sign carries meaning because a shared practice gives it a rule. I call this borrowed, or derived, content and distinguish it from original content. In the original case, the state's history and use fix what counts as right or wrong beyond the interpreter's rule. Originality means that independence, not self-creation; natural selection, learning, culture, or design may produce it, or may merely install our rule in a new mechanism. The dependence can vary by content and grain, and derived content remains real. The test asks where the standard comes from, not whether the state lies inside an animal or machine; it lets us grant content to a part without pretending that it speaks or that the whole system sees, thinks, understands, or feels.

A parrot can say "fire" without warning anyone of one, and a road sign can direct traffic only within a practice that gives its marks a rule. Some things mean *on loan*. Philosophers call this **derived** intentionality and contrast it with **original, or non-derived, content**: content whose correctness standard is not exhausted by an interpreting practice.

*Original* here means neither novelty nor self-creation. Language models can compose sentences no one has written before, and inherited vocabulary does not bar a learner from acquiring world-answerable uses its teachers never fixed. The question asks what now fixes the standard that makes a state accurate or mistaken, not whether the output appeared before or the system made itself. Novelty belongs to the product; originality concerns the relation among history, present use, world, and standard.

Across thermostats, animal nervous systems, trained mechanisms, inscriptions, and whole minded systems, the question stays the same: what fixes the standard by which the state gets things right?

### 14.1 The Thermostat Objection: Registration Without Perception

The thermostat Chapter 13 left on the wall has a second job. Early functionalists liked it because it made the mystery of mind look solderable: mentality as a sort of complicated thermostatting. The example still tempts the AI conversation.

Calling both a thermostat and a trained model representational leaves a question open: what kind of representation has each earned? The thermostat can carry derived content without perceiving the room. A learned model may do much more than relay a signal. Its richer organization deserves examination, but complexity alone does not tell us whether it perceives anything.

Tyler Burge saw why most cleanly. His *Origins of Objectivity* runs six hundred pages on essentially one question: what distinguishes a system that merely *registers* its environment from one that genuinely *represents* it? Burge treats the distinction as real, principled, and developmental. Birds, frogs, and infants cross it; thermometers, weather vanes, and detection circuits do not. The line matters for how we talk about machines.[1]

*Sensory registration* occurs whenever a physical system's inner state varies systematically with some feature of its surroundings. A thermometer's mercury column varies with temperature. A photodiode's voltage varies with light. A frog's retinal ganglion cells fire when a small dark spot moves across the visual field. These registrations carry what Fred Dretske called *natural information* about the environment[2] — the kind Chapter 13 showed comes cheap. Correlation alone has not established perceptual representation.

*Perceptual representation* requires a further distinction: between the stimulation reaching the sensors and the condition of the world beyond them. Burge calls that achievement objectification. In its own processing, a system must treat the proximal energy hitting its sensors as evidence about something *beyond* them. That distal condition could stay the same while the sensors change, or change while the sensors stay the same. This tests perception, not every form of content: a meaningful road sign need not perceive what it directs us toward.

### 14.1.1 Constancy as the Mark of Perception

The diagnostic signature of the achievement, Burge argues, is perceptual constancy. Constancies let the visual system hold a feature of the distal world fixed across wildly varying proximal stimulation. A white piece of paper in shadow reflects less light to your eye than a black piece in sunlight. Yet you perceive the first as white and the second as black. Color constancy tells you about the *paper's reflective properties*, not the *photons currently arriving*. A cat ten feet away projects a smaller retinal image than a cat at three feet, yet you perceive the same cat at two distances. Size constancy delivers the cat, not the angular size of its projection. Alison Springle puts the point cleanly: perceptual systems exhibiting constancy achieve a stable perception of actual macro-level physical properties of distal stimuli out of varying proximal stimulation. Burge uses constancy as the mark of perception proper: perception begins where the system starts solving for what's *out there* rather than merely responding to what's *coming in*.[3]

Does the thermostat distinguish its own bimetallic strip from the temperature of the room? It does not. Within the device, the strip exhausts the question; the thermostat solves for no distal fact and bridges no proximal-distal gap. Blow warm air directly onto the strip while the room stays cold, and it will report the room as warm. No level of its processing distinguishes the sensor from what the sensor is *of*. The thermostat instantiates a causal relation that we find useful to describe in representational terms.

John Searle made this last point, and made it with characteristic bluntness. In *The Mystery of Consciousness* and in the older "Is the Brain a Digital Computer?", Searle distinguishes between *intrinsic* intentionality, the genuine aboutness of a thought or perception, and *observer-relative* intentionality, which we attribute when a system merely admits of a useful interpretation.[4] His free-mapping argument gets that far: the same physical pattern can be assigned one syntax,

another, or none if interpretation alone does the work. It does not show that selection and use can never privilege one content over its rivals. The simple thermostat remains the easy case. Its strip controls the heating within an arrangement we built to regulate room temperature. That gives the state real derived content within our regulatory practice, without making the device a perceiver or establishing an original correctness standard.

### 14.1.2 The Frog's Subpersonal Machinery

The frog makes the more interesting case, because frogs and toads possess sensory systems sophisticated enough to make the question serious. When a frog's tongue flicks at a passing fly, what is the visual system doing? I find this the most instructive case in the chapter, precisely because the answer stays so hard to pin down. The classic ethological work — Lettvin and colleagues on the frog, J.-P. Ewert on the toad — is often read as suggesting something humbling: that the prey-capture circuitry responds to *small moving dark contours* rather than to flies as such.[5] That deflationary reading is contested, and the chapter does not need it to win. Ewert's own worm/anti-worm results show the circuitry responding selectively to a configuration, not a bare moving speck. Burge treats frogs' distance constancies as evidence that they perceive distal bodies. Whether they represent those bodies as flies, a natural kind, remains a different question. Ruth Millikan anchors representation in the evolutionary function of just such mechanisms.[6] Burge asks what the animal perceptually objectifies; teleosemantics asks how the frog's history and consumer economy determine what an internal state can get right or wrong *for the frog*. The circuit need not become a miniature subject. Its producer-consumer history may give its states subpersonal content, while the animal's wider organization determines what perceptual and cognitive capacities those states help realize. The next chapter isolates a producer-consumer organization without assuming that this circuit alone achieves objectification. It earns the thinner representational claim without denying perception to the frog.

Suppose a signal carries direction and distance to several consumers: one turns the animal, another aims its tongue. A blocked strike need not make the direction wrong. A lucky catch need not make it right. If the animal turns off target despite catching food, correction of the shared signal can reveal what its history established it to track. Chapter 15 develops that comparison without assuming that every frog actually has this organization.

Burge takes frogs to perceive distal bodies through constancies; he doubts that they represent flies as a natural kind. Our isolated circuit comparison assumes neither achievement. It asks what fixes a signal's content, while the animal's perceptual organization answers the further question of objectification. Burge's and Millikan's thresholds remain distinct even when an actual animal clears both.

### 14.1.3 Corpus Invariance and Perceptual Constancy

Bring this back to large language models trained and used only within a corpus, without a perceptual channel to a scene. Their tokens arrive as public signs produced by minds that solved their own proximal-to-distal problem. A model finds invariances across paraphrase, language, and coreference, but these are invariances of a corpus, not evidence of constancies that hold a perceived property fixed through changing stimulation. This corpus-only organization has not

established Burgean perceptual objectification, though its mechanisms may carry thinner content. Textual encoding alone settles nothing: sensor readings could also arrive as character strings, while a larger coupled system recovers distal properties across changing stimulation.

The reply is that humans also pattern-match on inputs. But human inputs are *proximal*, and the visual system uses them to recover distal properties stable across changes in stimulation — properties it can misperceive and whose detection was shaped by engagement with a physical environment. Byrne and Hilbert make the point about color: appearance tracks a distal surface property more closely than the photons currently arriving.[7] Without that separation of sensing from sensed, there is registration rather than perception.

The simple thermostat fails this distal test. The frog's circuit can carry subpersonal content within an animal that perceives. Corpus invariances alone do not establish the further perceptual achievement for a model. Chapter 15 asks how selection and causal use can give a mechanism correctness conditions without granting it every capacity of the containing system.

### 14.2 Derived vs. Original Intentionality: Borrowed Meaning

A well-trained language model produces a sentence about a maple tree. The sentence parses, predicts, and pleases. Did the model say something about a maple tree?

A growing line of argument answers yes — not by claiming the model perceives leaves or understands botany, but by locating content at a different grain. Its outputs belong to a public linguistic practice whose token-types have histories and uses. Its internal mechanisms have also been shaped by training. Those facts explain different achievements, and they should not be rolled together. The practice supplies borrowed public meaning; training supplies an acquired causal organization. Neither fact by itself makes the whole model a source of meaning.

Jumbly Grindrod's "Large Language Models and Linguistic Intentionality" works from Gareth Evans on naming practices and Ruth Millikan on teleosemantics.[8] Pointing out that no one has shown the model to be conscious misses his target. Grindrod sets sentience and consciousness aside and argues for *linguistic* intentionality that requires no mind on the speaker's end. On his account, public meaning lives in the practice, and any device that successfully participates in the practice bears it.

Words really do have public lives. "Water" refers to $H_2O$ whether the speaker can recite the formula; lineage matters. Millikan built her account to capture this: a token gets its content from the cooperative history that selected it.[9]

#### 14.2.1 Three Questions That Keep the Achievements Apart

The trouble starts when a meaningful output-token, a content-bearing mechanism, and a system that understands something get treated as one achievement.

Three questions keep them apart:

1. First, what fixes the content: a public practice, a training history, individual learning, or world-involving consumer use?

2. Second, at what level does the representational explanation operate: token, mechanism, circuit, model, or whole agent?
3. Third, how far does the content reach: a linguistic context, a distal kind, or a particular thing in the world?

No answer on one axis settles the other two. A token can carry public meaning while its producer lacks understanding; a mechanism can represent linguistic contexts while the larger system refers to no particular maple; and a content-bearing component need not become a believer.

### 14.2.2 What Training Selects

Millikan's mechanism has two halves. Producers make signs; consumers take them up; selection happens because consumer uptake helps determine which producer-tokens persist.[10] Artificial training can reproduce that pattern. The question is whether a proposed mapping helps explain the history or merely supplies an observer's gloss.

Fintan Mallory develops the strongest case for a genuine semantic result. On his consumer-based teleosemantics, word-embedding mechanisms acquire original content about linguistic contexts because training selects them to guide context-sensitive prediction.[11] He identifies a producer, a consumer, parameter changes under loss, and a contrast between successful and unsuccessful outputs. On the account used here, that described relation warrants original content about word-types and their contextual distributions when it obtains. Specific implementations and transfer to new consumers still require their own evidence.

Mallory supplies more than causal influence. His output is checked against word contexts found in the data, and the mismatch changes both producer and consumer. That gives the proposed content a worldly target. The remaining question asks what fixes the learned state's accuracy when that target and an assigned verdict part company.

Use the same comparison for a word model and a leak detector:

1. Specify the target and the grain. A distribution of word contexts differs from the things those words name. A coolant leak differs from the maintenance class assigned to it.
2. Trace how the proposed mapping shaped the mechanism and how a consumer uses it. Both sides of a comparison must have selection, correction, and consumer use. A trained system versus an inert chart proves too little.
3. Separate target, verdict, and payoff. Hold the target fixed while a grader changes its answer or a bonus hides failure. Then vary the target while holding those verdicts fixed. Ask which mismatch the established correction route registers and which state it retunes.
4. Follow that correction into consumer use. Does the revised state track the target across new cases, or only the assigned answers? Check the history as well as the response. A broken feedback channel can corrupt an established representation without changing what it represents.

Consider two context learners. Both receive the same word tokens, form embeddings, learn from error, and use the embeddings to predict nearby words. Their ordinary answers match. In one training history, target entries come from counts of the sampled contexts. In the other, they come from a schedule of approved answers that normally agrees with those counts. Now let the schedule

favor *engine* beside *cat* while the sampled contexts still favor *pet*. Keep the word key and context window fixed. A count-based mismatch can retune the first learner even if the displayed score praises its old answer. Correction in the second follows the schedule. We have described different sources of correction, not simply counted how often each learner succeeds.

That difference supports distinct readings. The first history supports a claim about actual contexts. The second supports the approved schedule; our further reading of its answer as a claim about actual contexts needs a separate basis. It might even have original content about the schedule. Failure at one grain cannot settle every other one. Nor can a single corrupt update erase an earlier history of tracking contexts. The comparison concerns what established the mapping, not which signal won on Tuesday.

Human practices determine what counts as a word and help choose the corpus. Once those are fixed, a judge cannot make *engine* common beside *cat* by approving the prediction. A child can learn that fact about language just as a machine can. Independence from a verdict does not require independence from the people who made the target.

Mallory identifies actual distributions as the source of training error. His account supplies the relevant dependence for the described algorithm: comparison with sampled contexts shapes the producer and consumer. That warrants the first reading at this limited grain. He does not report the matched experiment just proposed, but a fresh experiment supplies no missing semantic ingredient when historical and mechanistic evidence already establishes the relation. The comparison tests doubtful cases; it does not suspend the concession. A given implementation or a new consumer still requires its own evidence, and neither engineered loss nor linguistic ancestry counts against originality.

Relabeling demands a different control. Rename indices and adjust the matrices and word key together, preserving all links between actual words and contexts. The notation changes, not the candidate mapping. Switch which actual word feeds an index without that adjustment, and the input relation changes. Alter token boundaries or the context window, and the target may change too. None of these changes isolates an interpreter's verdict while leaving the content-fixing relations intact.

The concession remains narrow in reach. Hold the corpus fixed while the maple outside changes, and training on that corpus does not register the change. That limits a claim about the maple, not a claim about distributions in the corpus. The corpus belongs to the world too. Human uses of "maple" acquired wider address through practices answerable to maples. Statistical traces of those practices do not by themselves single out the tree outside my window.

Grindrod's public-practice argument presses at another level. He argues that the model can consume a naming practice in Evans's sense and inherit reference without the producer's demonstrative grip.[12] Grant that type-level inheritance: the practice fixes a public reference, and the model's token descends from it. What remains unearned by corpus training alone is occasion-level reference to *this* maple in a shared world. A consumer can inherit such a route through testimony and correction without going to look. We still need to establish that route for this use. Content about word distributions, even if original, does not supply it by itself.

### 14.2.3 Mechanism Content and Whole-System Understanding

Nothing in that concession makes an embedding a miniature speaker. Neural states can represent edges, motion, or color without believing anything; a frog's prey circuit can represent prey without becoming a second frog. The same distinction applies to an artificial mechanism. Content at one level does not automatically become belief, understanding, or reference by the larger system that contains it. A gearbox transmits torque without driving to the grocery store. Philosophy occasionally needs the same permission to distinguish a contribution from the achievement it serves.

Whole-system attribution depends on how the larger organization uses the content: integration across learning and inference, conflict resolution, continuing correction by what the state concerns, perceptual presentation, and flexible control. Such conditions may justify attributing a capacity to a whole agent. They do not create the component state's content after the fact.

The prayer analogy belongs here.[13] A transcription can preserve the words of a prayer without worshipping. Likewise, a mechanism can carry causally active content about textual distributions without the whole model understanding, asserting, or caring about any of it. That attributional limit holds whether the content remains derived or earns independence.

## 14.3 Dependence Without a Semantic First Mover

Calling content *derived* does not demand one uncaused first meaning. Every representational system has a causal pedigree. The contrast concerns explanatory dependence: does an interpreting practice exhaust the correctness standard, or do the system's world-involving producer-consumer history and present use help fix it?

Evolution, learning, culture, and training can all help fix this relation. Both context learners inherit a task. Both learn through error. Their histories differ in what drives that change: actual word occurrences or an approved-answer schedule. The source of their words settles neither case. Mallory's described producer-consumer relation warrants the limited word-context concession; the content of other implementations or transferred embeddings depends on their history and use.

Dennett offers the strongest reason to discard the distinction. On his real-pattern view, we earn literal talk of beliefs through robust, economical prediction. The pattern must hold across new cases and changes in conditions. No semantic spark or first interpreter needs to enter the story.[14] The book agrees that intentionality can emerge by degrees and that objective patterns need no ghost behind them. It asks one further question. Can actual contexts correct a prediction despite an approving verdict, or does the relevant history track the approved schedule alone? Dennett may treat that contrast as one more feature of the pattern. Nothing here forces him to adopt the original/derived vocabulary. The terms name a dependence that can vary while learning, useful prediction, and real content occur on both sides.

### Chapter Summary

**The claim.** Derived content depends on an interpreting practice; original content is world-answerable in a way not exhausted by one. Originality is compatible with causal and semantic inheritance, can vary by content and grain, and never means self-creation. Selection, learning,

and artificial training may establish it when their history fixes a standard rather than merely internalizing one.

**Where the argument stands.** A system can carry content without perceiving a distal object. The matched learning test asks what drives correction: the target, an assigned verdict, or payoff. Mallory's described relation warrants original word-context content when it obtains; other implementations and transferred uses still require evidence. None of this turns a circuit into a perceiver or a model into a speaker who understands.

**What's left open.** Integration, perceptual mode, attention, the distributed consciousness profile, and continuing world traffic may govern which wider system believes, perceives, understands, acts for reasons, or is conscious. They must earn those attributions directly rather than serving as hidden semantic conditions. The original/derived distinction diagnoses one kind of dependence; it does not rank every kind of mind.

**The hand-off.** A correctness standard can become world-answerable without any system making itself. The next chapter asks how producer-consumer history and a correspondence rule can establish that independence — and how much remains before content becomes a reason in a mind.

**Footnotes**

**1.** Tyler Burge, *Origins of Objectivity* (Oxford: Oxford University Press, 2010), esp. chs. 8–9. Burge's long argument distinguishes *sensory registration* — a state's covarying systematically with an environmental feature — from *perceptual representation*, which requires that the system itself treat proximal stimulation as evidence about a distal world it could be right or wrong about. The distinction is comparative and developmental: Burge sets the threshold of representation low enough to include many animals and human infants, but well above thermometers, weather vanes, and detection circuits. His disjunction-problem discussion, 323–324, specifically attributes distance constancies and distal-body perception to frogs while questioning perceptual representation of flies as a natural kind. The isolated circuit comparison in the text does not withdraw that animal-level attribution.

**2.** Fred Dretske, *Knowledge and the Flow of Information* (Cambridge, MA: MIT Press, 1981), esp. chs. 3 and 7–8. Dretske's *natural information* is a nomic-correlational notion, and the dependency runs from signal to source: a signal carries the information that a source is a certain way when, under the relevant conditions, the signal's occurring as it does is nomically sufficient for the source's being that way (equivalently, the conditional probability of the source's being that way, given the signal, is one). The step from carrying information to *representing* requires something further — for Dretske, the use of a semantic structure in shaping behavior — which is exactly where covariation and representation come apart.

**3.** Alison Springle, "Perception, Representation, Realism, and Function," *Philosophy of Science* 86, no. 5 (2019): 1202–1213, esp. sec. 2, develops the thesis that perceptual constancy delivers a stable perception of macro-level distal properties out of varying proximal stimulation; the main text paraphrases that thesis rather than quoting it. Burge makes constancy the criterial mark of perception proper (*Origins of Objectivity*, 2010, esp. ch. 9; see note 1): the constancies are the

mechanisms by which a system represents an objective, distally located property rather than merely registering proximal stimulation.

**4.** John Searle, "Is the Brain a Digital Computer?" *Proceedings and Addresses of the American Philosophical Association* 64, no. 3 (1990): 21–37, esp. 25–29; and *The Mystery of Consciousness* (New York: New York Review Books, 1997). The distinction between *intrinsic* (or *original*) intentionality and *observer-relative* (or *as-if*) intentionality is Searle's. His slogan "syntax is not intrinsic to physics" exposes the freedom of an interpretation that maps formal states onto arbitrary physical patterns. Chapter 22 accepts that objection to unconstrained maps while rejecting Searle's stronger conclusion: a physical mechanism can realize computational counterfactuals observer-independently. The semantic question remains separate; intrinsic computation does not, by itself, fix content.

**5.** Karen Neander, *A Mark of the Mental: In Defense of Informational Teleosemantics* (Cambridge, MA: MIT Press, 2017), esp. ch. 5. The prey-capture case — descending from Lettvin et al. on the frog and J.-P. Ewert on the toad — is the standard battleground for theories of content. It is worth flagging that Neander herself does not use it deflationarily: she argues the relevant states are low-level *perceptual representations* of visible features (a small, dark, moving thing), not mere registrations, as part of her solution to the indeterminacy problem. The main text needs less than either the "moving dark contour" deflation or Neander's own reading — only that a mechanism can carry consumer-imposed correctness conditions without clearing Burge's objectification threshold. Ewert's worm/anti-worm discrimination cautions in any case against reading the circuitry as a bare moving-speck detector.

**6.** Ruth Millikan (b. 1933), the American philosopher who built teleosemantics into a systematic theory of content — work the Royal Swedish Academy recognized with the 2017 Rolf Schock Prize — in *Language, Thought, and Other Biological Categories: New Foundations for Realism* (Cambridge, MA: MIT Press, 1984), esp. chs. 1–2; compactly restated in "Biosemantics," *Journal of Philosophy* 86 (1989): 281–297, esp. 284–88. Millikan grounds the content of a state in the *proper function* its ancestors performed — what the state was selected for — so that a mechanism can misrepresent precisely because it can fail to perform the function that fixes its content. The account is developed at length in Chapter 15.

**7.** The point belongs to the representationalist and reflectance-physicalist literature on color — Michael Tye, *Consciousness, Color, and Content* (Cambridge, MA: MIT Press, 2000), esp. ch. 7, 147–51; Alex Byrne and David R. Hilbert, "Color Realism and Color Science," *Behavioral and Brain Sciences* 26 (2003): 3–21, esp. 7–12. Color appearance tracks a distal surface property far more closely than it tracks the proximal photon flux currently reaching the eye — which is why color constancy holds the perceived color fixed as illumination changes.

**8.** Jumbly Grindrod, "Large Language Models and Linguistic Intentionality," *Synthese* 204:71 (2024). Grindrod sets aside general intelligence, sentience, and consciousness and recovers meaningful use at a different level: the linguistic practice itself. Drawing on Gareth Evans's distinction between *producers* and *consumers* of a naming practice (Evans, *The Varieties of Reference*, ed. John McDowell [Oxford: Clarendon, 1982], ch. 11), and on Millikan's teleosemantic account

of conventional signs, Grindrod argues that an LLM can stand in the consumer role even though it lacks the demonstrative-recognitional capacities Evans required of producers. The argument is the strongest version of the LLM-meaning move currently in print; the response developed here grants inherited public type-level meaning while denying that inheritance alone establishes occasion-level reference or whole-model understanding.

**9.** Ruth Garrett Millikan, *Language, Thought, and Other Biological Categories: New Foundations for Realism* (Cambridge, MA: MIT Press, 1984), chs. 1–2; and "Biosemantics," *Journal of Philosophy* 86 (1989): 281–297. Millikan's "proper function" is the function a trait has in virtue of the evolutionary or learning history that selected for its predecessors. Applied to signs, the proper function of a token-type is what its ancestors did that made it persist. The water-as-$H_2O$ case is most directly developed in Hilary Putnam, "The Meaning of 'Meaning,'" in *Mind, Language and Reality: Philosophical Papers, Volume 2* (Cambridge: Cambridge University Press, 1975), 215–271; the convergence between Putnam's causal-historical externalism and Millikan's teleosemantics is one of the more underappreciated agreements in twentieth-century semantics.

**10.** The producer-consumer asymmetry distinguishes Millikan's account from a purely informational starting point, not from a Dretskean theory confined to correlation. Dretske, *Knowledge and the Flow of Information* (Cambridge, MA: MIT Press, 1981), ch. 8, esp. 192–195, explains how learning establishes an information-carrying role that subsequent tokens can fail to fulfill. Information alone does not establish misrepresentation; his account introduces the additional functional history. Millikan locates the content-fixing work in consumers and the historically Normal conditions for their proper performance ("Biosemantics," sec. II, 284–290). The bee-dance example is Millikan's own (*Language, Thought, and Other Biological Categories*, 96–98). Peter Godfrey-Smith, "Mental Representation, Naturalism, and Teleosemantics," in *Teleosemantics: New Philosophical Essays*, ed. Graham Macdonald and David Papineau (Oxford: Oxford University Press, 2006), 42–68, provides the cleanest critical overview.

**11.** Fintan Mallory, "Teleosemantics for Neural Word Embeddings," *Mind & Language* (online 2026), https://doi.org/10.1111/mila.70037, esp. secs. 3–5.1. Mallory argues that word2vec-style embeddings possess genuine, original rather than derived content under a consumer-based teleosemantics. Section 5 treats selection through parameter alteration under loss minimization, distinguishing it from literal survival of unchanged weights. Section 5.1 grounds the probability-distribution interpretation in training and consumer use, not softmax mathematics alone. It distinguishes learned embeddings from their externally keyed one-hot inputs and explicitly treats linguistic distributions as worldly targets. His note 8 already discusses switched encoding vectors. A coordinated permutation may preserve the word-to-context mapping as a reparameterization; changing the external pairing instead changes a candidate content-fixing relation. Neither operation by itself establishes derivation. The chapter accepts original word-type/context-distribution content for the producer-consumer relation Mallory describes, conditional on its obtaining. Specific implementations and transfer to new consumers require their own evidence; a fresh matched experiment is not a missing semantic ingredient for the described relation. The ancestry of the vocabulary alone decides neither, and lack of demonstrative reference to a maple does not answer Mallory's proposed grain. No ownership condition resolves that dispute. The matched

comparison in the text is a proposed evidential control, not an experiment Mallory reports or a necessary capacity for online self-correction. Its teleosemantic basis is Millikan, "Biosemantics," *Journal of Philosophy* 86, no. 6 (1989): 281–297, sec. II, 284–290: selected and learned functions, historically Normal conditions, correspondence rules, and the distinction between proper success and accidental payoff. Mallory's author manuscript, sec. 4, pp. 7–8, and sec. 5.1, pp. 11–13, identifies the actual-data comparison; that is positive evidence of the relevant dependence, not merely numerical geometry. Separately varying corpus contexts, supplied target entries, and the reported score would test which relation explains the learned mapping. Equivalent historical-mechanistic evidence can support the same conclusion without a new intervention. The test neither requires a second owner nor makes cross-task flexibility a condition of all content.

**12.** Grindrod, "Large Language Models and Linguistic Intentionality," sec. 4. Millikan's framework explains content through selection-relevant producer-consumer feedback. Grindrod reads that framework as supporting the model's participation in a naming practice. The response developed here grants inherited public type-level meaning while denying that inheritance alone establishes occasion-level reference or whole-model understanding. On Millikan's consumer emphasis, see "What Has Natural Information to Do with Intentional Representation?" *Royal Institute of Philosophy Supplement* 49 (2001): 105–125. For an intralinguistic route to model meaning, see Vladimír Havlík, "Meaning and Understanding in Large Language Models," *Synthese* 205, art. 9 (2025).

**13.** The vehicle/content distinction at work here echoes Tim Crane's deployment of it in "Is There a Perceptual Relation?", in *Perceptual Experience*, ed. Tamar Szabó Gendler and John Hawthorne (Oxford: Oxford University Press, 2006), 126–146, and connects to the system-level challenge in Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3 (1980): 417–424. The prayer-transcription analogy is mine. Its limited point is attributional: a component may carry or transform genuine content without making the larger system a worshipper, speaker, or understanding subject.

**14.** Daniel C. Dennett, "Real Patterns," *Journal of Philosophy* 88, no. 1 (1991): 27–51, esp. 32–34 and 42–44; and *The Intentional Stance* (Cambridge, MA: MIT Press, 1987), chs. 1–2. Dennett's texts defend realism about predictively indispensable patterns and treat intentional attribution as answerable to more than an interpreter's whim. The formulation in the main text — especially its emphasis on novelty and counterfactual robustness — reconstructs the evidential lesson for the present argument rather than quoting a criterion Dennett states in those words.

## Chapter 15: Teleosemantics — Where Meaning Comes From

> *"He will starve to death surrounded by food if it is not moving."*
>
> — Lettvin, Maturana, McCulloch & Pitts, "What the Frog's Eye Tells the Frog's Brain"

CHAPTER OVERVIEW

How can meaning arise in one physical world? I argue that when some parts make signals and other parts use them, their shared history can fix what those signals should track and how they can go wrong. Teleosemantics names this account: learned or selected use can make a state answerable to the world in a way not exhausted by an interpreting practice. A frog's prey circuit shows how content can begin below thought; Swampman tests the role of a past, while pain's badness and mood remain beyond the explanation achieved here. The same test applies to machines, and content in one working part does not prove the larger system sees, thinks, reasons, or feels. Training can make borrowed language guide a mechanism from within, but that alone does not give it original meaning, understanding, or felt life.

A physicalist owes an account of how a physical state can acquire a correctness standard not exhausted by what an interpreter assigns it. The account need not locate a first meaning without a cause. The most developed attempt runs through biology: *teleosemantics*, from the Greek *telos*, purpose or end. Meaning is read from what a state is *for*. A frog's eye may represent a fly, airborne prey, or something nutritive at a location because of the work its signals do in the frog's evolved economy. Which description the signal earns is part of the problem, not a premise. Teleosemantics has its strength and its limit at the same point: function can constrain content without always fixing its exact grain.

### 15.1 Why Causal Grounding Is Not Enough

Start with the most direct answer: causal connection. A mental state represents water because water reliably causes it and it, in turn, causes water-seeking behavior. On this proposal, a representation gets its content from what it tracks along a causal chain from object to state.

Fred Dretske developed this idea most rigorously. On his account, a signal carries information that an environmental condition obtains when, given the relevant background, its occurrence guarantees that condition under the laws of nature. The guarantee runs from signal to source: water causing a signal does not suffice if other things produce the same signal too. Learning then recruits the connection to a use. A belief becomes a belief about water once the system has learned to form that state when water is present and use it to guide behavior.[1]

But causal connection alone does not suffice. A calibrated thermometer tracks temperature precisely without acquiring a mental state about it. Reliability does not by itself generate aboutness. Dretske therefore gave learning history the work of selecting content. Millikan went deeper.

She proposed biological function, specifically the function a producer-consumer organization acquired through evolutionary or developmental history.[2] Causal sensitivity and selected use together help determine what a state can get wrong. Return to Chapter 14′s frog. Nothing here reverses Burge's verdict that its detection circuit may fail to perceptually objectify a stable distal fly.[3] Teleosemantics asks a thinner question: can the signal represent what its consumers need and therefore misrepresent a BB? The claim must earn a prey-directed correctness condition; its exact grain remains unsettled.

### 15.2 The Frog That Couldn't Be Wrong

> So: the frog flicks its tongue, and this time it catches a small black BB someone fired across the pond. It swallows, looks satisfied, and waits for the next. The biologist writes *the frog mistook the BB for food*; the philosopher, watching from a folding chair, writes *did it, though?*

This is a paradigm of *occasional* misrepresentation: was this tokening accurate? The easy gloss says the state means *fly* rather than *fly-or-BB* because ancestral uses succeeded when flies, not BBs, were present. The example earns less. It may establish only *moving nutritive prey at this location*, with *fly* as our convenient label. The orthodox story excludes the BB branch only if pellets played no role in stabilizing the mechanism and satisfy none of the normal conditions under which its consumers succeed. Taxonomic grain does not come free. The harder case concerns *systematic* error: a whole perceptual apparatus reliably presenting the world in some respect as it is not.

That story is teleosemantics. A representation gets its content from the proper functions of the mechanisms that produce and consume it.[4] The frog's signal acquired a prey-directed role because tokenings in relevant ancestral conditions guided tongue strikes and feeding in ways that explain why this producer-consumer arrangement persisted. Its neurons know nothing about flies. Neither does selection. The history matters because it explains why this mechanism exists with this role rather than another.

The account extracts aboutness from causation, organization, and time, with no outside interpreter. Catching something nutritive may explain persistence where catching something black does not. Selection narrows the candidates; organization and use must do the remaining work.

Systematic error does not refute teleosemantics. A changed environment can make a historically selected mechanism repeatedly misrepresent while its correctness standard stays fixed. The account must explain that possibility. Selection history, task functions, tracking relations, and consumer use together constrain what the states represent.

Selection supplies equipment, task, and a history from which failure can be explained. It has not yet supplied a signal that *says* something. What turns failing into falsehood? And at what level should the representation be attributed?

### 15.3 From Failing to Falsehood

Put the thermostat and the frog beside each other. Both have jobs; both can fail them. The easy contrast says that the thermostat merely relays while the frog represents. A sophisticated thermostat spoils that answer. It may have graded sensors, several downstream policies, predictive control, and a long history of design revisions. Complexity, variation, and feedback do not draw the semantic line. The job must be stated without *indicates*, *accuracy*, or *content*. That is the job description challenge.[5]

Try a constructed comparison before stating the job:

> Build two pond controllers matched in detector sensitivity, a signal carrying direction and distance, feeding function, selection history relevant to the strike, and ordinary tongue-strike success. In the **Control** system, each coordinate pair feeds one tongue trajectory. In the **Map** system, the same pair also feeds orienting, obstacle avoidance, and deferred pursuit; consumers recombine it with body position and barriers. Block the straight strike. Control stalls. Map preserves the target variable and changes the route. Bias the detector and let a lucky deflection keep delivering prey. Map's orienting consumer now stops off target; that cross-consumer mismatch retunes the detector even while feeding continues. Jam the tongue and the relation may remain undisturbed despite failure.

The systems match in ordinary feeding, not in everything they would do under intervention. Their additional consumers and histories of use differ. That difference gives us something to investigate; it does not prove that adding parts creates meaning. Map illustrates three features of the proposed job:

1. A producer generates a family of states whose internal differences vary systematically with independently changing distal conditions.
2. Several consumers preserve those differences while selecting different responses.
3. The same state variable remains usable when the task, body position, or available route changes; it combines with separately changing inputs to produce responses never individually selected; and feedback retunes the producer when the state–world mapping drifts, even where some other part of the system rescues the final outcome.

Ward develops this decouplability test for directives. Its descriptive partner asks whether one distal-condition-sensitive variable can guide several possible behaviors rather than merely trigger one.

That description contains no semantic premise, yet it predicts a contrast. A relay tied to one response dies with the link. A family variable available across consumers transfers, recombines, and can be corrected at the mapping while downstream success stays fixed. These tests identify what the state does before deciding what to call it.

Millikan's correspondence rule names the structure that unifies those counterfactuals.[6] Each producer variant maps onto a distal condition, and that mapping figures as a Normal condition for the proper performance of several consumers — "Normal" here means historically explanatory, not statistically common. History privileges the rule that explains why these consumers preserve and exploit the same differences. A token's satisfaction condition comes from the world-side condition linked to its variant by that historically privileged rule. Nothing semantic entered the job description; mappings that fail its intervention tests never reach this step.

Correctness therefore cannot collapse into payoff. The mapped condition may obtain while the tongue jams, or fail to obtain while a lucky ricochet puts food in the frog's mouth. The rule also projects to variants no ancestor produced. Selection explains why this mapping governs the signal family; the world determines whether a present token satisfies it.

Jerry Fodor puts the sharpest pressure on the content's grain. A frog whose state means *fly* and one whose state means *fly-or-any-small-dark-speck* fare alike in an ancestral pond where the only specks were flies. Fitness alone cannot choose between them. The question is no longer whether the frog can fail, but what would make *fly* the thing it can fail to detect.

The consumer side of the loop constrains the answer without dictating a philosopher's favorite label.[7] Start with the task function whose performance causally explains why the producer-consumer mechanism stabilized. Then ask which tracking relation figures directly in the mechanism's performance of that task and how the organism's wider suite of capacities divides the work. If one circuit targets flying insects while different circuits guide the frog toward worms and other food, *fly at this location* may provide the most revealing explanation. If the same circuit and the same downstream use treat every small moving nutritive item alike, *moving prey* may exhaust the content. Taxonomy does not outrank the animal's organization.

This also answers the gruesome disjunction without treating absence from the ancestral pond as magic. *Fly-or-speck* is excluded only if the speck branch figures in no normal condition of consumer success, lacks the fly branch's nomological and counterfactual support, and improves no causal explanation of the system. Mere historical non-occurrence does not exclude it. *Edible moving thing*, *prey*, and *fly* pose the harder empirical choice. The supported content is the property, or structured family of properties, that makes a difference to performance across actual and relevant counterfactual tasks. If two candidates survive every consumer-relevant intervention equally well, content remains indeterminate at that grain. A frog need not carry a sharper concept than its life can use.[8]

Chapter 13 distinguished having a correctness condition from fixing its grain. A tie between *prey here* and *fly here* leaves the BB token false on either reading. A tie between *nutritive moving prey* and *small dark moving thing* matters more, since the BB counts as an error only on the first. Intervention can break the tie: hold darkness fixed and vary nutritiveness, then reverse the test. Which correlation makes a difference to consumer success is a counterfactual fact about the animal, not our description of it.[9] Where no asymmetry exists, the account accepts indeterminacy.

A tie that survives every consumer-relevant intervention leaves the account unable to choose between those content descriptions. With the complete property presentation and its mode held

fixed, choosing one description rather than another adds no phenomenal contrast on Chapter 6′s identity claim. Mechanism-level content supplies the semantic ingredient; integration and poising answer the separate questions of the state's conscious status and bearer.

### 15.3.1 Radical Enactivist Objection

Hutto and Myin press two claims. Much basic cognition does not need content: catching, reaching, navigating, and prey capture may consist in historically shaped organism–world coordination. And teleosemantics cannot create content from covariance plus proper function. Those facts explain sensitivity and malfunction, they argue, but never a condition of satisfaction. Calling the normal relation a correspondence rule merely gives coping a semantic title.[10] They are right about the one-link controller. It needs no representation, and selection does not care about truth.

### 15.3.2 Reply: What the Correspondence Rule Explains

Return to Map's lucky catch. Feeding continues while the orienting error retunes the shared signal. A description in terms of food obtained alone misses that correction; a richer contentless description might capture it.

Map's portability, recombination, and mapping-targeted correction distinguish it from direct control and satisfy Ward's job description for representational explanation. A radical enactivist can still describe the same common variable as part of a richer control organization. A nonsemantic description of that organization does not by itself remove content's explanatory role. The rival must show that the history-privileged mapping adds no unifying explanation of portability, recombination, and correction.

The remaining question asks what best explains the invariant. Several changing behaviors remain answerable to one distal condition; history privileges that relation; and whether the condition obtains can come apart from success in any response. I take that organized relation to constitute a satisfaction condition because it unifies portability, recombination, and correction under one rule. Hutto and Myin can call the same organization teleosemiotic rather than semantic. The physical account no longer borrows a semantic premise; the disagreement now concerns whether its history-privileged rule warrants the semantic verdict.

The frog does not discharge Chapter 7′s debt about pain and mood. Apply the same test to that chapter's proposed pain content: *a threat at this bodily site, bad for the subject*. A history of location-sensitive protection could fix the bodily address. The relevant condition concerns the threatened tissue, not whether withdrawal happens to succeed. Protection can fail despite an accurately located threat; a reflex or another person can prevent damage despite a mistaken location. Phantom pain then preserves an established bodily address without the corresponding limb. This explains how the address could be wrong, not how much the patient suffers.

The evaluative dimension has not passed the same test. Suppose two accounts predict the same protection of an aching ankle. One says the state presents a threat as bad; the other says a neutral threat signal prompts aversion. Distinguishing harmful from harmless changes, or giving one signal priority, does not choose between them. Even an accurate warning may tell us why to protect the ankle without telling us why undergoing the warning itself warrants relief. Calling

the state *bad-for-the-subject* would borrow the semantic description we need to earn. This chapter does not supply a non-circular naturalization of that dimension.

Mood leaves a similar gap. Bodily depletion and a day with few manageable options could help constrain Chapter 7′s body-and-field proposal. But imagine the kitchen still seems bleak after rest, with help available and the day's tasks well within reach. Does the mood misrepresent those conditions, present an unbound affective property, or cause other perceptions to change? The candidate targets do not yet decide. The achieved account reaches the perceptual mappings worked through here and a proposal for bodily address. It does not yet naturalize pain's badness, mood, or our grounds for ending an experience itself. I retain the universal identity thesis; I cannot count these unfinished applications as further explanatory successes.

### 15.3.3 Content Before Agency

Nothing in that verdict requires a free chooser. A frog's subpersonal mechanism may carry a correctness condition though the frog cannot compare a mortgage with a daughter's last year at home. A human deliberator may use words inherited from parents, teachers, and a public practice. Semantic inheritance does not disqualify the reason, and freedom does not retroactively mint the content.

The order runs the other way. First a world-involving causal-functional history fixes what a state gets right or wrong. Then integration may make that content available to a persisting subject across memory, comparison, correction, planning, and action. When differences in the represented consideration can bend conduct through that economy, the content functions as a reason. Chapter 18 calls the resulting graded capacity reasons-responsive agency. Moral answerability adds a further relation: a standing to receive a demand, understand it, and revise. Correctness, reason, agency, and answerability belong on four lines, not in one definition.

Making free will necessary for original content would close a circle around the theory. Free agency requires reasons; reasons require content; content would then require free agency. The circle explains nothing. The compatibilist lesson applies on both sides without making the sides identical. Causal inheritance does not prevent conduct from belonging to an agent when causes work through the agent's reasons-responsive organization. Semantic inheritance does not prevent a state from acquiring world-answerable content when learning and use establish a correction standard no longer exhausted by an external interpreter. Neither achievement escapes causation, and neither requires self-creation.

Free will does not mint content. It describes what an integrated subject can do with content: recognize considerations as reasons, weigh them, act through them, and revise when the world or another person answers back.

## 15.4 Representation and Its Level of Attribution

The last section earned a correspondence rule: a family of states varies in a way its consumers depend on, and the history of that dependence privileges some accuracy conditions over others. It does not yet tell us at what level to attribute the result — to a circuit, or to the animal the circuit serves.

The producer-consumer account already identifies a bearer at the level relevant to its explanation. A frog's visual circuit can carry content because variations in its states guide consumers whose proper functions depend on a mapping to prey-related conditions. The circuit need not become a tiny frog.[9b] Subpersonal representation concerns what a state does within a mechanism; personal-level belief concerns what the organized animal can perceive, learn, infer, and do by means of it.

The same split governs the artificial case. Mallory's analysis warrants original content about word-types and their contextual distributions when the described training and producer-consumer relation obtains. Actual linguistic data constrain the mapping; human ancestry alone does not make that standard merely borrowed. Chapter 14 gives the same narrow concession. Specific implementations and transfer to new consumers still need their own evidence. The semantic finding does not settle the level of attribution. An original representation would not put a speaker behind the vectors.

Integration can alter what we should attribute to the whole. A state used only by one reflexive circuit may support a thin subpersonal content. Coordinated across memory, learning, inference, conflict resolution, and action, the same content may help justify saying that an animal or machine believes or understands something. An external notebook or prosthesis can enter that organization when the coupling does the relevant cognitive work.

What none of it licenses is the leap from one content-bearing mechanism to a whole understanding subject. A representation need not own itself, and a subject cannot be assembled by changing the noun on every useful component.

The Control–Map comparison also connects intentionality with intelligence. A successful strike establishes success at that task. Map's ability to use the target variable for several responses, including a detour when the straight route closes, supports a claim of greater intelligence within that domain. The claim concerns flexibility across changed conditions, not how many successes we have counted. It requires no verdict about consciousness.

Intentionality addresses a different question: what the variable represents and what makes it accurate or mistaken. On the account defended here, its content helps explain Map's flexibility. Tracking the same prey position across changing routes lets consumers coordinate their responses without a separate signal for every movement. A contentless account remains a rival, as Section 15.3 acknowledged. The representational account earns its place only if the history-grounded mapping explains the transfer, recombination, and correction.

The same evidence can therefore bear on both questions without making them identical. Mapping-targeted correction tests accuracy; flexible rerouting tests problem-solving. Success can survive a mistaken target signal, and a jammed tongue can defeat an accurate one. A goal supplied by an engineer would not, by that fact, count against Map's intelligence or its states' content. Whether elementary adaptive control already counts as intelligence, and whether every such case requires content, remain open. Where representational content guides flexible problem-solving, intentionality contributes to the explanation of intelligence.

### 15.4.1 Michael Levin and Bioelectric Pattern Memory

A small flatworm makes the distinction less abstract. Cut a planarian into pieces and the surviving fragments can rebuild the missing parts. A fragment must produce a head at one end and a tail at the other. How does its tissue settle which goes where? At Tufts, developmental biologist Michael Levin and his colleagues study the electrical communication through which cells coordinate this work.

In experiments reported by Fallon Durant and colleagues, briefly disrupting channels between cells changed what some fragments rebuilt. They grew two heads. After the chemical washed out, further amputations still produced two-headed worms. Some treated worms initially looked normal, with one head and one tail, yet carried altered voltage patterns and could produce two heads after another cut. Present anatomy did not reveal the full regenerative disposition. A further intervention in membrane voltage restored stable one-headed regeneration in some animals.[9c]

Levin interprets such results as evidence of a rewritable *pattern memory*: a bodily state that helps specify the form to build, rather than merely recording the form already present. In his *Technological Approach to Mind Everywhere*, he places this work within a broader program that tests for cognition in unfamiliar systems by changing their conditions and measuring their responses. The proposal treats coordinated tissue repair as a candidate case of collective problem-solving, without waiting for a familiar brain before asking the question.[9d] Calling the activity intelligent would credit its adaptive problem-solving. Establishing what, if anything, its controlling states represent remains a further task.

I take the planarian work to supply an experimental candidate for the kind of representation this chapter describes. A retained physical state guides the production of an anatomy that does not yet exist. On the representational reading, its content concerns what to build, not what the worm currently looks like. A directive need not match the present world; it specifies a condition the ensuing work should bring about. Nor does anything inside the worm have to inspect a little picture of its future body. The cells respond to one another, and their responses change the body.

Hutto and Myin's objection still has teeth. A stable, rewritable controller need not represent anything; changing its setting changes what it does. The experiments establish causal control and persistence, but they do not by themselves show that a semantic explanation improves on a contentless one. Section 15.3 supplies the further questions: which differences in the retained state do downstream processes use, across which changes in injury, and what history makes the proposed mapping explanatory? Producing two heads cannot alone establish misrepresentation. If the state directs the construction of two heads and the tissue complies, it has met that directive, however abnormal the outcome.

Levin's work gives those questions a tractable biological setting. It supports the search for representation below the level of a whole thinking animal and shows how to investigate the candidate without adding an inner reader. It does not establish that regenerative tissue feels anything, or that phenomenal character consists in representational content. Those claims carry further burdens. The worm can help us investigate what a bodily state does without telling us what, if anything, undergoing it feels like.

### 15.5 What Embodiment Contributes

Descartes held that thought does not essentially require a body. The *res cogitans*, the thinking thing, could in principle exist disembodied. He also recognized the intimate union of mind and body in human pain, hunger, and thirst: a sailor merely observing damage to a ship would not feel the ship's injury. The disagreement concerns the mind's claimed independence, not whether bodily life shapes human experience.

This gets things backwards. Embodiment supplies much of what makes a mind possible. Here *embodiment* names a functional achievement, not a carbon clause. A biological body realizes it one way; an artificial or distributed system could realize it another if the relevant loops actually close.

1. Embodiment first supplies *sensorimotor traffic*: cycles of perception and action through a resistant environment in real time. Reach for a cup and vision guides the hand, proprioception registers the arm, touch registers contact, and movement changes what vision receives next. Perception and action form one loop, with the world running through it.
2. It may also supply *autonomy*, when that traffic helps maintain the organization doing the sensing and acting. An organism lives exposed: it can be depleted, damaged, or killed. Reciprocal self-maintenance may help explain why the organism counts as an individual agent and where its practical interests lie.[9a] It does not supply the content of its states.
3. Embodiment supplies *a history of encounters* that tunes sensitivities and fixes addresses. How you understand a forest path turns partly on paths that resisted and yielded, on falls, recoveries, and learned footing. A nonbiological system could acquire an equally consequential history through other sensors and effectors. What matters is not skin but traffic that selects among possible readings of its states.
4. Finally, embodiment supplies *a located point of view*: a here from which things can be near or far, reachable or blocked.[11] Not every later user of a name must stand where its referent once stood. I can refer to Napoleon because testimony and deference pass along a handle that someone in the chain acquired through contact. Chapter 16 asks what that chain must contain. For now, the locating work need occur somewhere in the practice, not under every speaker's feet. A functionally embodied system may inherit reference through a community and acquire new addresses through its own traffic. It cannot conjure a distal particular from internal form alone.

### 15.6 Swampman

> One objection cuts deeper, and it arrives dripping. Donald Davidson's Swampman is a lightning-struck duplicate assembled by cosmic accident: molecule for molecule yourself, his visual cortex and color circuitry matching yours to the last synapse, but with no history whatever.[12] At the first instant his organization and dispositions match yours. He lacks any etiological or learning history that could privilege

one content over another. Swampman can preserve himself and still lack teleosemantic content at $t = 0$.

Strict Millikan and strict Dretske make history necessary for content. This chapter accepts the price. Present contact can begin a causal history, but it cannot retroactively supply one. Swampman's first state may covary with a tomato and cause reaching, sorting, and eating. Until learning recruits such relations into functions, they remain causal dispositions rather than representations. His consequential traffic can begin the history from which content later emerges.

Duplication settles neither content nor identity. A manufactured copy of a selected artifact may inherit a function through its lineage; an accidental duplicate inherits none merely by matching the structure. Copying makes succession wonderfully ugly. Philosophers managed that well enough before the lightning arrived. What matters for content is causal provenance and producer-consumer history, not whether the resulting system repairs itself.

On the wide strong representationalism adopted here, Swampman lacks phenomenal character at $t = 0$. There is nothing it is like to be him then. Dretske accepted this historical cost.[13] The claim concerns more than a missing name for this tomato. Swampman lacks the history that, on this theory, gives his states any perceptual property content. In Chapter 6′s twin-tomato case the object changes while that content survives. Here the property layer itself lacks its historical basis.

The structural rival can grant the missing past and still count Swampman's present visual organization as enough for the same look. It need add no mental paint. Picture both duplicates turning toward a red tile: matched discriminations and reports supply no independent test of whether only one has experience. Our verdict follows from combining the history requirement with the wide identity; the example does not confirm either premise. Chapter 6′s blur and attention comparisons support the broader intentionalist account more strongly than this historical extension. I retain the initial absence verdict at that lower level of warrant. Millikan and Neander also resist the usual Swampman intuition; neither intuition nor its rejection decides the issue.[14]

The comparison with a trained language model now concerns history, not manufacture. Swampman lacks a content-fixing history. Mallory's described training and consumer use warrant original linguistic-context content at the mechanism's grain. The occurrences of those words in actual text belong to the world, even though public practice produced the words. That finding does not extend automatically to every text-trained model or a new consumer. Model-level understanding remains open. A history that establishes linguistic-context content does not by itself integrate it into belief, planning, reference to particulars, or a point of view. ### Chapter Summary

**The claim.** A family of signals, its correspondence rule, and its producer-consumer history can establish original content at the grain the explanatory organization supports when that history fixes a world-answerable standard not exhausted by an interpreting practice. This independence requires no semantic self-creation.

**Where the argument stands.** Selection, learned sensitivity, and consumer use explain how a state can get something wrong and how interventions can choose among candidate contents. A

frog's subpersonal circuitry may represent prey without becoming a separate subject. Mallory's described producer-consumer relation warrants original content about word-types and context distributions; other implementations and transferred uses still require evidence. Neither conclusion establishes whole-system understanding or distal reference beyond the history that fixes the content. The book retains its claim that Swampman at first has no experience. That claim has weaker support than the ordinary perceptual account. The content-fixing cases do not establish it independently.

**What's left open.** This chapter has not explained pain's badness, mood, or the grounds for relief in physical terms. They add no achieved success to the identity thesis. How content guides reasons, becomes conscious, and belongs to a subject also calls for separate accounts.

**The hand-off.** A distinctively human reach runs through words, which refer to a world that helps fix what they mean outside the skull. Chapter 16 turns to that reach.

**Footnotes**

**1.** Fred Dretske, *Knowledge and the Flow of Information* (Cambridge, MA: MIT Press, 1981), esp. ch. 3; and *Explaining Behavior: Reasons in a World of Causes* (Cambridge, MA: MIT Press, 1988), esp. chs. 3–4. Dretske's account develops from an information-theoretic foundation: a signal carries information about a state of the world when the signal's occurrence is nomically sufficient for the world's being in that state (under certain idealized conditions). The move from information-carrying to *representing* requires the additional step of explaining how information gets put to use — how the organism learns to exploit the information to guide behavior. This move involves the organism's learning history and is where the connection to teleosemantics begins to emerge.

**2.** Ruth Millikan, *Language, Thought, and Other Biological Categories* (Cambridge, MA: MIT Press, 1984), esp. chs. 1–2; and *White Queen Psychology and Other Essays for Alice* (Cambridge, MA: MIT Press, 1993). The core claim of teleosemantics is that the content of a representation is determined by its *proper function* — the function it has in virtue of its evolutionary or learning history — rather than by what it actually tracks on any given occasion. This explains misrepresentation: a detector misfires when it responds to a pellet because its proper function is to detect prey, not pellets, even though on this occasion it is causally responding to a pellet. On the frog: the neurophysiology is Lettvin, Maturana, McCulloch, and Pitts's (1959), and the philosophical use of the frog case in the disjunction debate is canonically Fodor's (see note 8); Millikan's own paradigm cases in *Language, Thought, and Other Biological Categories* and "Biosemantics" are the beaver's splash, the bee dance, the hoverfly, and the toad's bug-detector. For a useful critical overview, see Peter Godfrey-Smith, "Mental Representation, Naturalism, and Teleosemantics," in *Teleosemantics: New Philosophical Essays*, ed. Graham Macdonald and David Papineau (Oxford: Oxford University Press, 2006), 42–68.

**3.** Tyler Burge, *Origins of Objectivity* (Oxford: Oxford University Press, 2010), esp. chs. 8–9. Burge's constraint is that perceptual representation requires objectification — constancies that present distal properties across changing proximal stimulation. Nothing here contests that threshold for perception. The separate teleosemantic question concerns thinner, subpersonal content fixed by

producer-consumer organization; granting such content does not turn the circuit into a perceptual subject.

**4.** The thesis that content derives from selected function is Millikan's; see note 2 above for *Language, Thought, and Other Biological Categories* (1984) and "Biosemantics" (1989), and note 1 for Dretske's learning-based variant. The revised account adds no ownership condition. Selected producer-consumer function may fix what a mechanism can misrepresent when the proposed mapping, rather than an observer's gloss, helps explain the mechanism's stabilization and survives the relevant interventions.

**5.** William Ramsey, *Representation Reconsidered* (Cambridge: Cambridge University Press, 2007), ch. 1, poses the job description challenge: say what a state does *as a representation* rather than merely as a causal relay. Zina Ward, "Directive Representation and the Job Description Challenge," *Philosophy and the Mind Sciences* 7 (2026), doi:10.33735/phimisci.2026.12005, argues that a directive earns representational status when its specific contribution remains decouplable from any one behavior. Variation, a consumer-presupposed correspondence rule, and the history that makes that rule explanatory supply the present reply.

**6.** Ruth Millikan, *Language, Thought, and Other Biological Categories* (Cambridge, MA: MIT Press, 1984), 96–102; "Biosemantics," *Journal of Philosophy* 86 (1989): 281–297, at 283–290; and *Varieties of Meaning* (Cambridge, MA: MIT Press, 2004), 71–86. Every representation belongs to a set of interrelated representations a producer can send to a consumer, and it is Normal for consumer performance that the set map onto world conditions under a rule. The rule and its explanatory history, not a further owner, make the content non-derived. "Normal" means historically explanatory, not statistically usual.

**7.** Millikan's consumer emphasis appears in *Language, Thought, and Other Biological Categories*, chs. 1–2 and 6, and "Biosemantics." Nicholas Shea develops an explanatory version in *Representation in Cognitive Science* (Oxford: Oxford University Press, 2018), especially chs. 3, 4, and 6. Task functions inherit the grain of the causal explanations that stabilize them, and content depends on which correlations play an unmediated role in performance. Kevin J. Mitchell, "The Origins of Meaning: From Pragmatic Control Signals to Semantic Representations," *Philosophy and the Mind Sciences* 7 (2026), doi:10.33735/phimisci.2026.11674, traces how action-bound signals become usable across wider cognitive organization. Wider integration may warrant system-level attribution; it does not create the component content.

**8.** Fred Dretske, *Explaining Behavior: Reasons in a World of Causes* (Cambridge, MA: MIT Press, 1988), chs. 3–4, and *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995), ch. 1, where the merely *indicating* properties of a state (what it nomically correlates with) and its *representational* properties (what it has the function of indicating, acquired through a learning history that recruits the indicator to a use) can come apart. The "thickening" image follows Dretske's account of learned indicator-functions, but no naturalistic theory owes a distinction where rival contents remain explanatorily equivalent across every relevant counterfactual. The frog-and-fly disjunction in its influential form is Fodor's: see Jerry Fodor, *A Theory of Content and Other Essays* (Cambridge, MA: MIT Press, 1990), and, on the charge that his asymmetric dependence

presupposes the semantic and functional facts it was meant to ground, Peter Schulte and Karen Neander, "Teleological Theories of Mental Content," *Stanford Encyclopedia of Philosophy*, substantive revision 26 May 2022, secs. 4.1–4.2. The detailed fly, prey, and disjunction round appears in Appendix B.15.

**9.** The objection is that an appeal to which correlations play an *unmediated* role in explanation may track the interests of the explainer rather than a joint in the mechanism. Versions of it are pressed against Shea's framework by Todd Ganson, who argues the distinction carries no explanatory significance, and by Frances Egan, who is deflationary about the utility of representational explanation generally; both are surveyed in Schulte and Neander, "Teleological Theories of Mental Content," sec. 4, with Shea's own reply in *Representation in Cognitive Science*, ch. 6. The reply pressed in the body does not turn on the explainer at all, and it does not merely observe that the history happened. Which correlation makes a difference to consumer success with the others held fixed is itself a counterfactual fact about the mechanism, so the *selection among* historical facts that the objection targets is as objective as the facts selected.

**10.** Daniel D. Hutto and Erik Myin, *Radicalizing Enactivism: Basic Minds without Content* (Cambridge, MA: MIT Press, 2013), esp. chs. 3–4, argue that covariance and proper function never yield truth-conditional content. The disagreement rests entirely on Section 15.3′s correspondence rule. The rule specifies a satisfaction condition for each signal variant, including unused variants; accuracy can diverge from present success; and producer-consumer history makes the mapping explanatory rather than optional. Hutto and Myin may deny that this suffices. Adding a proprietor would not answer them, because it would identify a candidate subject without specifying what any state says.

**9b.** Teleosemantic accounts routinely locate producers and consumers below the organism without turning each subsystem into a subject. Nicholas Shea treats vehicles as constituent parts of a larger system in *Representation in Cognitive Science*, ch. 1; Peter Schulte and Karen Neander make the same allocation explicit in "Teleological Theories of Mental Content," *Stanford Encyclopedia of Philosophy*, revision 26 May 2022, sec. 2. Andy Clark and David Chalmers extend cognitive processes across reliably coupled resources in "The Extended Mind," *Analysis* 58, no. 1 (1998): 7–19. These cases support a distinction between mechanism-level content and the wider level at which belief, agency, or understanding should be attributed. Nicholas Shea's unity-of-purpose requirement addresses that later threshold ("Realist Representational Explanations of Agency Should Require Some Unity of Purpose," *Philosophy and the Mind Sciences* 7 [2026], doi:10.33735/phimisci.2026.12098). For the artificial mechanism case, see Fintan Mallory, "Teleosemantics for Neural Word Embeddings," *Mind & Language* (online 2026), doi:10.1111/mila.70037, esp. secs. 3–5.1. Mallory establishes differential retention, consumer use, and a privileged causal mapping.

**9c.** Fallon Durant et al., "Long-Term, Stochastic Editing of Regenerative Anatomy via Targeting Endogenous Bioelectric Gradients," *Biophysical Journal* 112, no. 10 (2017), doi:10.1016/j.bpj.2017.04.011, especially figs. 1 and 4 and the sections on cryptic planaria and bioelectric resetting. The initially single-headed, anatomically normal animals did not all later become double-headed: the cryptic state involved stochastic regeneration, whereas the double-headed state persisted across further amputations. Pharmacological hyperpolarization restored stable single-

headed regeneration in 34% of the double-headed animals tested (N = 102). The study supports a functional role for bioelectric state while leaving open contributions from other interacting epigenetic mechanisms. Persistence, reversibility, and dissociation from present anatomy support the pattern-memory interpretation; they do not by themselves establish semantic content or isolate its precise physical vehicle.

**9d.** Michael Levin, "Technological Approach to Mind Everywhere: An Experimentally-Grounded Framework for Understanding Diverse Bodies and Minds," *Frontiers in Systems Neuroscience* 16 (2022): 768201, doi:10.3389/fnsys.2022.768201, especially pp. 16–19 and fig. 7. Levin explicitly interprets the planarian case as counterfactual representation of target morphology. The teleosemantic assessment offered here is my application of this chapter's criteria, not a claim that Levin endorses Millikan's theory or this book's identity thesis. The distinction between an altered directive and a false description matters: deviation from species-typical anatomy does not independently identify a semantic error.

**9a.** Mark Bickhard grounds representation directly in self-maintenance in "Representational Content in Humans and Machines," *Journal of Experimental and Theoretical Artificial Intelligence* 5, no. 4 (1993): 285–333, and "The Interactivist Model," *Synthese* 166, no. 3 (2009): 547–591. Wayne Christensen and Cliff Hooker connect autonomy to anticipative learning in "An Interactivist-Constructivist Approach to Intelligence: Self-Directed Anticipative Learning," *Philosophical Psychology* 13, no. 1 (2000): 5–45. Alvaro Moreno and Matteo Mossio's *Biological Autonomy* (Springer, 2015) systematizes organizational closure. These remain important theories of autonomy and agency. Closure bears here on autonomy and agency; any semantic work must come from the producer-consumer mapping and its history.

**11.** The argument that embodiment is constitutive of cognition rather than incidental to it draws on a broad tradition in phenomenology and cognitive science. Key sources include Maurice Merleau-Ponty, *Phenomenology of Perception*, trans. Colin Smith (London: Routledge & Kegan Paul, 1962 [1945]), whose account of the lived body as the ground of perception anticipates much contemporary embodied cognition research; Francisco Varela, Evan Thompson, and Eleanor Rosch, *The Embodied Mind: Cognitive Science and Human Experience* (Cambridge, MA: MIT Press, 1991); and Andy Clark, *Being There: Putting Brain, Body, and World Together Again* (Cambridge, MA: MIT Press, 1997). The specifically philosophical argument about what embodiment contributes to semantic grounding is developed in Gareth Evans, *The Varieties of Reference*, ed. John McDowell (Oxford: Clarendon Press, 1982), esp. ch. 6, whose account of demonstrative thought and spatial cognition shows how perceptual and action capacities underlie the ability to think about particular objects. Descartes distinguishes mind and body while expressly affirming their union in sensation: *Meditations*, Meditation VI, in *The Philosophical Writings of Descartes*, vol. II, trans. John Cottingham, Robert Stoothoff, and Dugald Murdoch (Cambridge: Cambridge University Press, 1984). The separability and bodily-union passages were consulted in MIT's *Minds and Machines* course excerpt, "Sixth Meditation: the real distinction between mind and body," pp. 1–2 of that excerpt, not verified against the volume's original pagination. The objection in Section 15.5 concerns separable intellect, not a denial by Descartes that ordinary human sensation depends on bodily union.

**12.** Donald Davidson, "Knowing One's Own Mind," *Proceedings and Addresses of the American Philosophical Association* 60 (1987): 441–458, at 443–444, where the Swampman case is introduced — not as an objection to teleosemantics, which was not its target, but to press the point that Davidson's own externalism about content makes causal history essential to what a state means. Teleosemantics inherits the case because it makes a stronger and more specific historical claim.

**13.** The consequence is *phenomenal externalism*: if phenomenal character is representational content of the right kind, and content is fixed in part by world-involving relations, then phenomenal character fails to supervene on the intrinsic physical state of the subject's body taken alone — molecular duplicates can differ phenomenally. Fred Dretske accepted the commitment explicitly and defended its compatibility with physicalist supervenience, once the supervenience base is widened to include the world-involving facts that fix content: *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995), ch. 5, "Externalism and Supervenience," 123–168; and "Phenomenal Externalism or If Meanings Ain't in the Head, Where Are Qualia?", *Philosophical Issues* 7 (1996): 143–158. Alex Byrne and Michael Tye defend phenomenal externalism, not thereby this same historical test for Swampman, in "Qualia ain't in the Head," *Noûs* 40 (2006): 241–255. The physicalism at issue never promised intrinsic-duplicate invariance; it promised no mental difference without a physical difference, and a missing causal history is a physical difference. Ned Block presses the contrary intuition — that felt character must be narrow — as part of the mental-paint package; see "Mental Paint," in *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, ed. Martin Hahn and Bjørn Ramberg (Cambridge, MA: MIT Press, 2003), 165–200. Appendix B.9 sets out the pressure and the reply. Dretske's chapter 5, especially 144–151, argues against treating the internalist intuition as decisive; he explicitly distinguishes this defense from a proof of externalism. Byrne and Tye separate externalist intentionalism from particular reductive psychosemantics (opening, secs. 1.1.1 and 2). Tye's own treatment permits experiences in accidental replicas under suitable nonhistorical tracking conditions (*Consciousness, Color, and Content*, sec. 6.1, 120–122, and sec. 6.4, 136–140); agreement about phenomenal externalism must not be converted into agreement about the book's initial-absence verdict. The present verdict retains the lower warrant of the historical extension identified in Chapter 6.

**14.** Karen Neander, "Swampman Meets Swampcow," *Mind and Language* 11 (1996): 118–129, argues the intuition should be overturned, not accommodated; Ruth Millikan, "On Swampkinds," *Mind and Language* 11 (1996): 103–117, argues that a swamp creature falls under no real (historical) kind, so generalizations from humans to swampmen carry no support — a case her *Language, Thought, and Other Biological Categories* (cited in full at note 2 above) had anticipated with a swamp-replica of its own (93), three years before Davidson's. The dialectic, including the "real nature" strategies and Papineau's actual-world alternative, is surveyed in Schulte and Neander, "Teleological Theories of Mental Content," sec. 4.2; Appendix B.9 works through the case in full.

## Chapter 16: Language and the World

> *"Cut the pie any way you like, 'meanings' just ain't in the head!"*
>
> — Hilary Putnam, "The Meaning of 'Meaning,'" 144

---

CHAPTER OVERVIEW

The same words, **the window is open**, can tell you a fact, ask you to close it, or warn you of danger. I want to explain how a sentence gets its meaning, what a speaker does with it, and why understanding takes more than the right words. Grammar lets familiar parts form new thoughts, while shared use gives speakers standards they can learn, dispute, and correct. I argue that some uses must reach beyond signs through contact with things, other people's words, and acts that can go wrong, though many words name nothing. A machine may produce good English and show real skill within a field; whether it also warns, keeps a promise, or uses what it learns across one continuing life depends on more than the sentence we see.

### 16.1 One Sentence, Several Achievements

"The window is open."

Suppose I say it while a summer breeze lifts the curtain: a report. Say it as rain crosses the sill and you sit by the latch: probably a request. Say it after glass breaks downstairs: a warning.

The words, order, and window stay fixed. What I mean and do change.

The sentence presents a condition: the window stands open. My purpose may go further: I want you to close it. In speaking, I may also make a request for which you can reasonably ask why. Content, speaker meaning, and the act performed with the words already come apart.[1] You can understand the request perfectly and still refuse it.

Chapters 12–15 supplied the metasemantic background: form cannot select a duplicate, correlation lacks an error standard, and producer-consumer history may establish one. A frog can represent prey without words.

Language adds public signs, combinatorial structure, communicative purposes, acts, and social correction.

Calling all of this *use* leaves us something to explain. Use includes teaching, correction, perception, consequence, and action. Text prediction may reward astonishing competence while directly comparing outputs only with signs. Which wider relations make those signs about rain, arthritis, or this window?

My answer keeps use and reference together. Not every word names an object. Rather, some uses must answer to objects, events, conditions, or socially maintained kinds beyond relations among signs. Those routes determine the practice's empirical subject matter. Reference names this worldward dimension of use, not an added ingredient.

Control and Map supplied the nonsemantic test. Map preserved and corrected one target across consumers; Control only triggered a response. That world-directed organization, not a hidden mind, does the grounding work.

A greeting need not name anything, and a promise can fail without referring badly. Frege and Russell show why reference cannot mean attaching labels to every word.

### 16.2 Sense, Reference, and Empty Expressions

The simplest theory of meaning treats words as labels. *Cat* means cat because the word stands for cats. *London* means London because it names a city. The view offers a pleasing economy: find the object, attach the word, and send language on its way.

Two old puzzles spoil the arrangement.

Gottlob Frege spent his career trying to show how thought could preserve both objectivity and logical structure. In 1892 he asked why identifying *the morning star* with *the evening star* can teach someone something while identifying the morning star with itself cannot. Both expressions in his example refer to Venus. If reference exhausted meaning, the two identity statements should carry the same cognitive value. They plainly do not.[2]

Frege distinguished a sign's **reference**, what it stands for, from its **sense**, the way that referent gets presented. *The morning star* and *the evening star* reach the same planet by different routes. A person may grasp one route without knowing that it joins the other.

Sense also lets us understand an empty expression. *The largest prime number* presents a condition although no number satisfies it, and a fictional name can guide a reader without acquiring a flesh-and-blood bearer. Language can organize a thought before the world supplies an object.

Bertrand Russell wanted a different repair. In "On Denoting," he turned to the present King of France. The phrase looks like a name, but Russell analyzed it as contributing existence, uniqueness, and predication conditions without placing any king inside the proposition. *The present King of France is bald* says, roughly, that exactly one person now reigns as King of France and that this person is bald. The existence condition fails. The sentence comes out false without requiring a royal ghost to make it meaningful.[3]

Grammatical shape does not settle semantic function. A noun phrase may quantify rather than contribute an object; other expressions modify or connect. Discourse can fail to refer at one point while remaining answerable to a world at another.

The qualification locates reference at the level of a linguistic practice. Fregean sense, Russellian analysis, logical vocabulary, fiction, questions, and promises enrich that practice without making reference every word's job. None shows that a complete natural language with determinate empirical subject matter could float free of any world.

Consider a frightened child who says, "There is a monster under my bed." No monster answers, but the child has not produced noise. Stories and pictures give the word recognizable criteria. The sentence reaches toward the world and misses. Missing differs from never reaching.

The distinction becomes decisive in the artificial case. A language model's emitted words already belong to English. Their tokens inherit conventional, derived meaning from the practice that supplied the training corpus. We should not deny that meaning merely because the model did not invent English. Printed books do not invent it either. The remaining question concerns the system's own referential acts. Does *water* in this output merely reproduce a public word whose reference others fixed, or does a continuing artificial system itself use the word to track, correct, remember, and act upon water? The visible token cannot answer that question alone.

Words seldom travel alone. Language owes much of its power to how they combine.

### 16.3 How Sentences Build Meaning

A child learns a few thousand words and soon understands sentences nobody has previously uttered. You can read *Three purple elephants discussed mortgage rates beneath the moon* without consulting a phrasebook devoted to elephant finance. The sentence's novelty does not block comprehension. Familiar parts and familiar modes of combination let you work out an unfamiliar whole.

Philosophers and linguists call this feature **compositionality**: the meaning of a complex expression depends systematically on the meanings of its parts and on how those parts are arranged.[4] Word order matters because *the dog bit the postman* and *the postman bit the dog* contain the same vocabulary and report different disturbances. Structure makes finite language productive.

Compositionality explains a real achievement. It also tempts a familiar overreach. Once a formal system can represent syntax and calculate how parts combine, one may think semantics has emerged from the machinery. But a rule of combination preserves and transforms whatever semantic contributions its inputs already possess. It does not establish that *dog* concerns dogs, that *water* reaches $H_2O$, or that a speaker's claim answers to the world rather than to a sequence of marks.

The distinction here runs between a **semantic theory** and a **metasemantic theory**. A semantic theory tells us what expressions mean and how their meanings determine the contents of larger expressions. A metasemantic theory asks what facts make those expressions have those meanings in the first place.[5] The previous chapters supplied part of the metasemantic story: causal history, selection, learning, public practice, and producer-consumer use. Compositionality explains how content, once grounded, can travel through structure.

Donald Davidson gave this semantic ambition one of its most influential forms. Rather than treating meanings as another class of objects attached to sentences, he proposed a systematic theory that derives the truth conditions of each sentence from the semantic roles of its parts: "Snow is white" is true if and only if snow is white.[6]

Truth conditions do not exhaust language. Questions, commands, jokes, metaphors, and context-sensitive expressions require more. Davidson's point concerns systematicity: finite resources must explain how speakers grasp indefinitely many sentences.

This matters for language models because compositional competence counts as evidence of more than memorization. A model that handles new combinations, preserves dependencies across long

passages, and adapts an expression to a novel context has learned substantial regularities of language.

It does not yet answer the grounding question. Two systems can manipulate the same sentence structure while their words reach different things, as Twin Earth will make vivid. A third may preserve every formal relation while inheriting its semantic interpretation from the people whose language it models. Composition can carry meaning a long way. It cannot choose where the journey began or what lies at its destination.

### 16.4 Meaning in Use

Ludwig Wittgenstein approached ordinary words with the suspicion that philosophy had lost its way by leaving their use behind. In *Philosophical Investigations*, he gave up the search for one hidden essence and looked instead at the varied lives of words.

"For a large class of cases—though not for all—in which we employ the word 'meaning' it can be defined thus: the meaning of a word is its use in the language," Wittgenstein wrote.[7] The qualification matters. He asked us to inspect language-games: ordering, describing, guessing, joking, thanking. Speaking belongs to an activity, a form of life.

The shift breaks the private-label picture. Imagine someone who calls a recurring private sensation *S* and records each occurrence in a diary, relying on memory to guide the next entry. What would distinguish remembering correctly from merely feeling satisfied? If whatever seems right counts as right, correctness has collapsed into apparent correctness. A sign whose application answers only to what seems right at each use cannot supply its own standard.[8]

Nothing in that argument denies sensations, introspection, first-person authority, or talking to oneself. Solitude causes no difficulty. The difficulty comes from stipulating a use for which apparent correctness exhausts every possible standard of correctness.

Rules work through practices. Children learn words amid examples, correction, consequence, and shared attention. A carpenter who calls for a slab expects a slab, not a beam; another participant can correct the mistake, and the building can expose it. No monitor needs to watch every later application. Correct and incorrect already have a place in the learned activity.

A public standard does not turn consensus into truth. Many speakers and inherited classifications treated whales as fish. Public use made disagreement intelligible; anatomical and evolutionary evidence, working through the aims of biological classification, forced a revision. The community did not vote mammals into existence, and the world did not write the taxonomy for us.

Wittgenstein's investigation locates correctness within norm-governed activity; it does not offer a causal theory of reference. The synthesis I draw from it goes further. Meaning-as-use and reference reinforce each other because meaningful use already includes practices of identifying, testing, correcting, and acting upon things. A practice remains open to objects, events, and people that can prove its applications wrong.

Kripke, Putnam, and Burge add a different consideration without turning Wittgenstein into an externalist: which causal histories, environments, and social relations help determine the people,

substances, and kinds a practice concerns? The combination belongs to the account developed here. First we need to notice what speakers do with content once they have it.

### 16.5 What Speakers Do with Words

Return to the open window. The sentence says that the window stands open. That content remains available whether I report the fact, warn you about an intruder, or request that you close it. Sentence meaning constrains what I can reasonably mean. It does not finish the job.

H. P. Grice had an unusually attentive ear for the work ordinary conversation performs. He drew the decisive distinction between signs that naturally indicate something and speakers who mean something. Dark clouds can mean rain without intending an audience. Spots can mean measles. Communicative meaning differs. A speaker who says, "Take an umbrella," may intend a hearer to recognize that rain is coming and to form that belief partly by recognizing why the words were spoken.[9]

The account explains how speakers communicate more than their sentences literally say. Adapt one of Grice's examples: ask a colleague whether Smith makes a good philosopher and hear, "His handwriting is excellent." The sentence concerns handwriting. In the circumstances, the speaker probably means that Smith's philosophy inspires no comparable praise.[10]

Grice organized such reasoning around a Cooperative Principle and conversational maxims: give the exchange what it requires, avoid what you believe false, stay relevant, and speak clearly. Actual conversation departs from these maxims cheerfully and often. Implicature frequently arises when a speaker conspicuously flouts one while leaving cooperation in place. Irony would have a difficult career otherwise.

J. L. Austin widened the view. Philosophers had concentrated on statements and their truth. Austin pointed to utterances that perform actions: *I apologize*, *I promise*, *I name this ship*. Saying the words under the right conditions does not merely report an apology, promise, or naming. It carries one out.[11]

Austin distinguished three acts. A **locutionary** act produces meaningful words with a certain sense and reference. An **illocutionary** act does something in saying them: asserting, warning, promising, or requesting. **Perlocutionary** effects follow from saying them, as when speech persuades, alarms, or amuses. John Searle refined the structure by separating utterance acts, propositional acts, and illocutionary force. The same reference and predication can appear in an assertion, a question, or a command.[12]

Words and grammar contribute content. Context helps determine what gets said. A speaker uses the content with a communicative purpose and may perform an act that carries a force and a commitment. The hearer may infer still more.

Not every speech act answers to truth in the same way. Assertions commit speakers to how things stand. Directives aim to make conduct fit words. Promises place the speaker under an obligation to make later conduct fit what was said. Questions solicit an answer. Each has conditions under which it succeeds, misfires, or gets withdrawn.

The same gap reopens for artificial systems because emitting an assertion-shaped sentence and undertaking an assertion may come apart. A system can print *The bridge is safe*. Designers, users, or an institution may use that output to issue a warning or certify the bridge without making the mechanism a speaker on its own account. Has the system itself committed to the claim? Can evidence defeat the commitment? Will the correction govern later answers and plans? Which continuing organization stands behind the words tomorrow? Fluency supplies no automatic answers.

### 16.6 How Language Reaches the World

Wittgenstein's grammatical investigation has already refused the picture of meaning as private mental cargo. It neither asks nor answers the later metasemantic question: why does *water* reach $H_2O$ rather than the clear liquid on Twin Earth, or why does *arthritis* exclude an ailment in the thigh despite one speaker's confident usage?

Saul Kripke attacked the idea that every speaker carries a description that fixes a name's bearer. Most people who use *Feynman* cannot uniquely describe Richard Feynman. They inherit the name through a chain of communication reaching back through books, teachers, colleagues, and earlier speakers. Somewhere in that history, uses connect to the person. Later speakers can refer successfully while knowing very little.[13]

The chain involves more than causal descent. A later speaker normally intends to use the name with the reference it has in the practice. By the same condition, calling a dog *Feynman* can begin a new chain precisely because the speaker does not intend to preserve the physicist's name-bearing relation. That condition will matter when we ask whether a trained mechanism participates as a linguistic consumer or merely transmits tokens whose reference others fixed.

Chapter 6 visited Twin Earth to show why its lakes look like ours. Its original point concerned words. Hilary Putnam built Twin Earth with a mathematician's discipline: alter one variable and hold the rest fixed. On that planet, a liquid called *water* looks and behaves like ours but has a different chemical structure, XYZ. Imagine Oscar on Earth and his psychological counterpart on Twin Earth before either community knows chemistry. Their inner states match. Their terms do not reach the same substance. Oscar's practice developed around $H_2O$, while Twin Oscar's developed around XYZ. Their environments help fix what their words concern.[14]

Tyler Burge showed how the community enters the picture. His patient believes he has arthritis in his thigh, though arthritis affects joints. The patient has not changed the meaning of *arthritis* by misunderstanding it. Medical and linguistic practice supplies a standard against which his use comes out mistaken. A counterfactual community that used the word more broadly would give the same inner state a different content.[15]

Kripke supplies causal history, Putnam environmental anchoring, and Burge social individuation. Gareth Evans adds a useful division of labor. **Producers** maintain recognitional or informational contact with a referent. **Consumers** inherit the ability to use the term through deference and communication. Each of us consumes far more names than we produce. Reference does not require every speaker to meet every object. Consumer standing still involves participation in the practice: prior producers count as authoritative, and correction can alter later use.[16]

I take these externalist arguments to fit the account of use rather than replace it. That compatibility forms part of this book's synthesis, not a doctrine Wittgenstein stated. A public practice reaches the world through an ecology: producers encounter and investigate; consumers inherit, defer, and correct their use when that ecology pushes back.

### 16.6.1 The conceptual-role objection

A stronger challenge goes beyond empty names. A conceptual-role theorist argues that meaning can consist in the place an expression occupies within reasoning and use. We understand *justice* through the judgments it licenses, *wit* through contrasts and applications, and even a fictional king through the network of claims in which he appears. Language models can learn enough of that network to draw novel inferences. Why demand a further word-world relation? Steven Piantadosi and Felix Hill press this case directly for large language models: relations among a system's internal states may realize important aspects of meaning without reference. Robert Brandom, a contemporary American pragmatist, builds a broader inferentialism around what speakers commit themselves to, what claims they have reason to make, and which inferences depend on content rather than form. His account also includes entry moves from perception to claims and exit moves from claims to action, so it does not close the network around words alone.[17]

The challenge identifies something real. Inferential role explains important aspects of sense and competence. A narrow version claims that relations among internal states can fix subject matter by themselves. Imagine two practices whose token-to-token inferences and actual-world use profiles match. In Actual, speakers produce *red* through a detector whose history and organization make it track red surfaces. In Twin, the same token and internal transitions answer to a magnetic field. Redness and the field covary throughout the speakers' actual environment, so their ordinary uses match. Intervene to break the correlation by presenting a red surface without the field, however, and the different producer routes show themselves.

Brandom rejects representational reference as a primitive building block of meaning. His inferentialism instead includes reliable circumstances in which speakers apply a term, practical consequences of applying it, and social practices of commitment and correction. On his account, a practice comes to count as representing the world through that norm-governed activity rather than through a semantic atom added first. This gives the practice worldly friction in Brandom's own terms.

That broader answer satisfies the practice-level requirement rather than defeating it. Once reliable entry and exit moves count as part of an expression's role, the practice no longer closes around words alone. I need not claim that reference comes before use. I claim that meaningful use in a complete natural language must include some world-reaching routes that determine empirical subject matter. The Twin case targets the narrower claim that internal relations suffice.

A metasemantic question still remains for any world-facing role: what makes an entry move answer to redness rather than to the field? Producer history, consumer use, and learning constitute the different routes; interventions reveal which route continues to govern when the correlations come apart. If none privileges a target, the content remains indeterminate.

This establishes the practice-level requirement without assigning a referent to every word. A complete natural language must contain some routes by which its uses answer to nonlinguistic or socially maintained conditions beyond relations among signs. Fiction, mathematics, logical vocabulary, and other subpractices can take their semantic place within that wider language. What, if anything, fixes reference to abstract objects remains open. The present claim needs only that a complete language cannot close around its own signs while retaining determinate empirical subject matter.

### 16.6.2 The empty-expression objection

Fregean sense and Russellian descriptions may seem to show that meaning needs no reference. They show something narrower: an expression can fail to denote, or perform no referential job, while the larger practice remains answerable beyond its signs. *The present King of France* fails against a world in which France has no king.

The objection therefore changes the level of the requirement rather than defeating it. Empty expressions defeat a universal denotation requirement, not the practice-level referential condition. Their applications, implications, and conditions of satisfaction still belong to a language that answers beyond the signs.

### 16.6.3 The inherited-reference objection

The stronger AI objection grants externalism and turns it in favor of language models. Matthew Mandelkern and Tal Linzen argue that a model's words may inherit reference through the human uses recorded in its training corpus. If ordinary speakers can consume a name without meeting its bearer, why cannot a model inherit reference from the same chain?

Here the objection succeeds, up to a point. A generated token, the mechanism that produced it, or causally active internal states may inherit semantic or referential content through the public tokens and corpora used in training. A blanket denial would give human consumers a privilege the theory itself rejects.

I was once too quick to treat corpus descent as semantically inert. That was a mistake. Training can make public linguistic distinctions causally active inside a mechanism, and a deployed system may participate further through correction, tools, perception, memory, and action.

Evans's consumer still participates in a practice: producers remain authoritative, correction alters later use, and an intention to preserve the inherited reference helps keep the chain joined. A mechanism need not reproduce that intention in human form. Nor must a one-off token persist in order to inherit semantic reference. Persistence bears on the different question whether one continuing system refers, retracts claims, or undertakes commitments on its own account. The relevant evidence asks whether corrections survive, govern later inference or action, and keep that organization answerable to the practice's producers.

That evidence also identifies the bearer. A transient model, a persistent agent, and an embodied controller may use the same language mechanism while differing in which corrections survive, what observations constrain them, and whether later conduct answers to earlier words. The product label does not decide where the achievement belongs. The control organization does.

Mollo and Millière strengthen the externalist case for machines. Mediated causal information, selected functions, and world-sensitive training may ground content in model outputs or internal states. I accept that possibility without taking it to establish understanding, knowledge, or consciousness in the larger system.[18]

A brain envatted after a lifetime of worldly engagement can retain reference established before the vat. Putnam's always-envatted brain lacks the ordinary causal route to brains and vats. An interactive system, by contrast, may form a genuine route to stable features of its physical or simulated environment. Where rival interpretations remain tied, reference may remain indeterminate. The medium can change without making inner traffic sufficient. History and use do the work.[19]

### 16.7 When a Machine Uses Language

Readers who take current language models to understand have good reason to resist the phrase *mere fluency*. Contemporary systems do more than repeat memorized sentences. They preserve grammatical structure across novel combinations, adapt to context, exploit implication, translate, summarize, correct local errors, and shift register. Those capacities provide real evidence of linguistic competence. The models' internal states can carry causally active derived content inherited from public language. Calling all of this "just statistics" replaces analysis with tone.

Fluency gets a system into the argument. It does not decide it.

Imagine three systems saying *The window is open*. The first replays a recording when a light sensor flips. The second uses a learned classifier to track the sash and generate a report. The third belongs to a household agent that notices rain, checks the latch, warns you, and closes the window when authorized. The recording carries meaningful words its designers can use to warn; the classifier adds tracking; the agent coordinates the report with a purpose and an action.

Let the household agent use that same classifier, and let it mistake a reflection for a gap. You show the agent that the sash has closed and ask it to check the latch before warning you again. In the stronger version of the case, that correction changes its stored expectation, its next check, and its later report even after the conversation ends. Merely replacing the current sentence would achieve less. The correction gives us something to test: does the same organization learn what went wrong and use the lesson, or does someone outside it have to carry the lesson from one component to another?

That difference supplies evidence of understanding within this household task. It does not require every adult human capacity, and it does not establish consciousness. Whether the agent itself warns, rather than simply delivering a warning delegated by someone else, also depends on its communicative purpose and the commitments its later conduct sustains.

Six claims must therefore remain separate.

1. **The output instantiates meaningful English expression types.** Their conventional meanings come from the public practice; visible shape alone would not suffice.
2. **The token or generating mechanism may carry semantic content.** Corpus-mediated history can transmit derived content or reference to an output token and make related

distinctions causally active in internal states. Corpus descent alone does not establish original intentionality or show that the larger system refers on its own account.

3. **The system displays linguistic competence.** It reliably exploits syntactic, semantic, contextual, and pragmatic regularities across new cases. This capacity does not presuppose understanding.
4. **The system may mean something by its words.** Gricean speaker meaning requires a communicative purpose available for audience recognition. A designer or user may also communicate through a mechanism that lacks a purpose of its own.
5. **The system may perform a speech act.** An utterance can function as an assertion, promise, warning, or request through convention or delegated authority. A system that undertakes the act on its own account must also bear the corresponding commitment. Persistence helps us attribute such commitments across time; it does not define illocutionary force.
6. **The system may understand to some degree.** Bounded understanding begins when several content-guided capacities constrain one another and remain answerable to correction. Breadth across memory, perception, inference, learning, planning, and action can make that understanding more integrated and world-involving. No single transcript settles how far it extends or which bearer has earned the attribution.

Havlík presses a different challenge: a sufficiently rich linguistic fragment may sustain minimal semantic content and understanding through its contextual organization. I do not think relations within the fragment alone fix a distal subject matter, but they can support a real bounded capacity. Burge's arthritis patient supplies a human precedent for separating possession of content from complete mastery without establishing that any particular model crosses the threshold. Partial understanding can still count as understanding.[20]

In this book I will sometimes put quotation marks around a language model's "understanding." They mark a limited, domain-specific capacity rather than a counterfeit one. The bearer may amount only to a running mechanism or deployed product during a bounded exchange: enough organization to interpret a question, draw relevant inferences, correct an error, and recombine what training supplied. Integrated personal understanding requires a further showing that memory, perception, purposes, and correction belong to one persisting, world-answerable organization. Chapter 17 takes up that question. Bounded understanding alone establishes neither original intentionality nor phenomenal consciousness.

These achievements do not form a ladder, and none may impersonate the others.

None of the six settles whether the system is conscious or whether a persisting subject stands behind the words. Those questions require their own evidence.

### 16.7.1 The other-minds objection

A defender of that attribution can press a fair objection. We do not inspect another person's causal history and internal organization before accepting that person's words. We hear fluent, context-sensitive speech and attribute understanding. Why demand more from a machine?

The history behind the charge deserves respect. AI debate has often treated yesterday's impossibility as today's irrelevant trick. The answer cannot consist in raising one bar each time a system clears the last one.

Human speech sometimes reaches us only as a transcript. Even then, our interpretation draws on extensive background evidence about the kind of continuing life that produced it: a life with perception, memory, correction, plans, and actions available in principle to further observation. We do not inspect that organization each time. The background makes fluency a reliable cue; when the cue comes apart from that background, its evidential weight falls. None of this provides certainty, and Chapter 19 rejects analogy as the universal basis of our knowledge of other minds, not as a possible aid. The point concerns evidence, not delivery medium.

The same standard should govern artificial systems. Biology receives no automatic privilege. When an artificial system preserves commitments, learns from correction, coordinates language with perception and action, and carries those gains through one continuing organization, the evidence rises. Where a product exposes only fluent output while persistence, correction, and world-involving control remain absent or scattered across components coordinated only by external users, the transcript underdetermines the verdict. Equal treatment requires the same distinctions, not identical assumptions about unlike cases.

**Chapter Summary**

**The claim.** A language with determinate empirical subject matter requires reference at the level of the whole practice, though many words name nothing. Some uses must reach beyond signs to objects, events, and shared kinds. Without such routes, the practice cannot fix what its claims concern.

**Where the argument stands.** Reference neither exhausts meaning nor comes before use. Form and inference alone cannot fix what a language concerns. Public rules, speaker aims, speech acts, causal history, the world, and social standards each do different work, so a system can show language skill without thereby earning every other success.

**What's left open.** Fluent artificial systems may produce good English and tokens or inner states with inherited content. They may show real skill and, in a limited sense, "understand" within a field. We must still ask which bearer has original rather than derived content, refers or communicates on its own account, takes on commitments, and brings its content-guided skills together.

**The hand-off.** The next chapter turns from meaning to the one who means: a subject assembled through memory, language, and social life.

**Footnotes**

**1.** The distinctions among expression type, utterance, propositional content, speaker meaning, and illocutionary force cross the conventional semantics/pragmatics border. See Kepa Korta and John Perry, "Pragmatics," *Stanford Encyclopedia of Philosophy*, substantive revision 28 May 2024, secs. 1–2; Mitchell Green, "Speech Acts," *Stanford Encyclopedia of Philosophy*, substantive revision 24 September 2020, secs. 1–2. Green's force/content discussion also explains why one proposi-

tional content can occur in assertions, questions, and directives without reducing the differences among those acts to differences in reference.

**2.** Gottlob Frege, "Über Sinn und Bedeutung," *Zeitschrift für Philosophie und philosophische Kritik*, n.s. 100, no. 1 (1892): 25–50; Max Black, trans., "On Sense and Reference," in *Translations from the Philosophical Writings of Gottlob Frege*, ed. Peter Geach and Max Black (Oxford: Blackwell, 1952), 56–78. The informative-identity argument begins from the difference in cognitive value between $a = a$ and a true $a = b$, then distinguishes *Sinn* from *Bedeutung*. The consulted translation uses *morning star* and *evening star*; the later Hesperus/Phosphorus relabeling does not appear in Frege's 1892 text.

**3.** Bertrand Russell, "On Denoting," *Mind* 14, no. 56 (1905): 479–493. Russell's contextual analysis treats a definite description as contributing existence, uniqueness, and predication conditions to the proposition expressed rather than standing for a peculiar denoting entity. Later work divides sharply over whether ordinary definite descriptions carry Russellian quantificational semantics, Strawsonian presuppositions, or referential uses in Donnellan's sense. For the contemporary map, see Peter Ludlow, "Descriptions," *Stanford Encyclopedia of Philosophy*, substantive revision 21 September 2022.

**4.** Bryan Pickel and Zoltán Gendler Szabó, "Compositionality," *Stanford Encyclopedia of Philosophy*, substantive revision 3 November 2025. The principle requires specification of the relevant syntax, semantic values, and determination relation; without those auxiliary constraints, compositionality can become formally trivial. Conversational implicatures and many context-dependent contributions do not compose in the same direct sense as literal semantic content.

**5.** Jeff Speaks, "Theories of Meaning," *Stanford Encyclopedia of Philosophy*, substantive revision 31 July 2024, distinguishes semantic theories—which assign semantic properties to expressions—from foundational or metasemantic theories, which explain the facts in virtue of which expressions possess those properties. The book's Chapters 12–15 operate chiefly at the metasemantic level; the present chapter adds semantic structure, speaker meaning, and force without treating those as rival answers to one question.

**6.** Donald Davidson, "Truth and Meaning," *Synthese* 17 (1967): 304–323, at 311 for the *Snow is white* test case. Davidson proposed that an empirically adequate Tarskian truth theory could serve as a theory of meaning for a natural language if constrained to yield interpretively acceptable theorems from finite resources. The proposal concerns systematic semantic competence; it does not claim that truth conditions alone supply a complete theory of language use, force, or the metasemantic grounding of primitive expressions. See Jeff Malpas, "Donald Davidson," *Stanford Encyclopedia of Philosophy*, substantive revision 28 April 2023, secs. 2–3.

**7.** Ludwig Wittgenstein, *Philosophical Investigations*, trans. G. E. M. Anscombe, 3rd ed. (Oxford: Basil Blackwell, 1967; English-text reprint, 1972), secs. 23 and 43. Section 43 carefully limits the meaning-as-use formula to "a large class of cases—though not for all"; Section 23 emphasizes the open-ended multiplicity of language-games and places speaking within activity and a form of life. The chapter therefore uses *meaning as use* as a diagnostic against private mental cargo, not as a one-clause analysis of every semantic phenomenon.

**8.** Wittgenstein, *Philosophical Investigations*, secs. 201–202, 240–242, and 243–258. The rule-following considerations distinguish grasping a rule in practice from indefinitely interpreting one formulation through another. The private-diary case at sec. 258 attacks a proposed criterion whose seeming-correct exhausts correctness; secs. 241–242 distinguish agreement in language and judgments from agreement that votes propositions true. The claim that public practice, causal history, environment, and social relations together determine content forms the synthesis developed in this chapter, not a theory attributed to Wittgenstein.

**9.** H. P. Grice, "Meaning," *The Philosophical Review* 66, no. 3 (1957): 377–388. Grice distinguishes natural meaning from nonnatural or communicative meaning and develops the latter through an utterer's intention to induce a response by means of the audience's recognition of that intention. Later Gricean accounts revise the exact reflexive-intention conditions, especially for conventionalized, collective, covert, or noncooperative communication; the main text uses the durable distinction rather than treating Grice's first analysis as final.

**10.** H. P. Grice, "Logic and Conversation," in *Syntax and Semantics*, vol. 3, *Speech Acts*, ed. Peter Cole and Jerry L. Morgan (New York: Academic Press, 1975), 41–58. Grice distinguishes what is said from conversational implicature and characterizes the latter through rational reconstruction under the Cooperative Principle and maxims of quantity, quality, relation, and manner. The Smith/handwriting exchange in the main text adapts Grice's testimonial for "Mr. X" rather than quoting his example. Cancellability, calculability, and nondetachability remain useful diagnostics, not exceptionless definitions of implicature.

**11.** J. L. Austin, *How to Do Things with Words*, ed. J. O. Urmson (Oxford: Clarendon Press, 1962). Austin begins from explicit performatives and the distinction between felicitous and infelicitous acts, then weakens the performative/constative dichotomy in favor of a general theory of locutionary, illocutionary, and perlocutionary acts. Statements themselves count as illocutionary performances and remain assessable along dimensions beyond bare truth or falsity, including uptake, authority, and circumstance.

**12.** John R. Searle, *Speech Acts: An Essay in the Philosophy of Language* (Cambridge: Cambridge University Press, 1969), chap. 2. Searle distinguishes utterance acts, propositional acts of reference and predication, and illocutionary acts such as asserting, questioning, and commanding. The same propositional content can occur under different forces; conversely, repetition of the same sentence type across contexts need not preserve propositional content or force.

**13.** Saul A. Kripke, *Naming and Necessity* (Cambridge, MA: Harvard University Press, 1980), especially the second lecture, 91–97. Kripke's causal-historical sketch replaces the requirement that each competent user associate a uniquely identifying description with a name. The transmission condition includes an intention to use the name with the reference current in the practice. The dog named *Feynman* in the main text adapts Kripke's Napoleon/pet-aardvark example. The sketch leaves substantial questions about failed baptisms, speaker reference, deference, and reference change; it supplies a world-involving constraint rather than a complete algorithm.

**14.** Hilary Putnam, "The Meaning of 'Meaning,'" in *Language, Mind, and Knowledge*, ed. Keith Gunderson, *Minnesota Studies in the Philosophy of Science* 7 (Minneapolis: University of Minnesota

Press, 1975), 131–193. Putnam's Twin Earth argument directly establishes an extension difference between Earthian *water* and Twin-Earthian *water*. His broader account adds syntactic and semantic markers, stereotype, extension, and the division of linguistic labor. The chapter therefore does not slide from difference of extension to a complete theory of meaning without argument. A slide explainer of the thought experiment appears at mindmatterandmeaning.com/explainers/twin-earth.

**15.** Tyler Burge, "Individualism and the Mental," *Midwest Studies in Philosophy* 4 (1979): 73–121. Burge's arthritis case extends externalist pressure from natural kinds to socially individuated concepts: the subject can possess and misuse a concept whose correct application depends partly on communal linguistic practice. The conclusion concerns content individuation, not an infallible community or a denial of first-person authority.

**16.** Gareth Evans, *The Varieties of Reference*, ed. John McDowell (Oxford: Clarendon Press, 1982), chap. 11. Evans distinguishes information-linked producers from consumers who inherit a name through communication. Consumer standing does not require direct acquaintance with the bearer; it requires participation in a practice whose information can ultimately be maintained and corrected. The distinction becomes important for artificial systems because corpus descent may transmit public reference while leaving system-level correction and attribution open.

**17.** Robert B. Brandom, *Making It Explicit: Reasoning, Representing, and Discursive Commitment* (Cambridge, MA: Harvard University Press, 1994), especially chap. 4 on language-entry and language-exit moves; Steven T. Piantadosi and Felix Hill, "Meaning without reference in large language models," arXiv:2208.02957 (2022). Brandom gives inferential role its strongest systematic form while retaining reliable world-facing entry moves; Piantadosi and Hill press conceptual-role meaning directly in the machine case. The twin-practice argument targets the claim that relations among internal states suffice. Brandom's broader role already includes entry differences; Chapters 12–15′s producer-consumer and intervention tests address the further metasemantic question of what makes one entry route answer to one target rather than another, without appealing to any criterion of practical stake.

**18.** Matthew Mandelkern and Tal Linzen, "Do Language Models' Words Refer?" *Computational Linguistics* 50, no. 3 (2024): 1191–1200, https://doi.org/10.1162/coli_a_00522, argue that externalist causal chains may extend through training corpora to language-model tokens. Dimitri Coelho Mollo and Raphaël Millière, "The Vector Grounding Problem," *Philosophy and the Mind Sciences* 7, no. 1 (2026): 1–28, https://doi.org/10.33735/phimisci.2026.12307, distinguish several kinds of grounding and defend routes to referential grounding for trained systems. Both lines block a manufacture-based denial of reference; neither makes reference sufficient for integrated understanding or phenomenal consciousness.

**19.** Hilary Putnam, *Reason, Truth and History* (Cambridge: Cambridge University Press, 1981), chap. 1, stages the always-envatted brain and argues that its terms cannot refer to brains and vats in the ordinary way because the required causal relations never formed. The recently-envatted contrast follows from standard externalist continuity rather than from Putnam's compact presentation. Reference to stable virtual objects or feed structures, and indeterminacy where rival

mappings remain tied, are extensions of the argument made here rather than claims attributed to Putnam.

**20.** Burge's arthritis case shows that a speaker may possess a socially individuated concept while misunderstanding part of its extension; see Burge, "Individualism and the Mental," 77–79. For the machine case, Vladimír Havlík, "Meaning and Understanding in Large Language Models," *Synthese* 205 (2025): article 9, https://doi.org/10.1007/s11229-024-04878-4, argues that minimal semantic content and semantic fragmentism support attributing natural-language understanding to LLMs without making awareness a condition of understanding. I accept the pressure against treating the achievement as mere simulation, but not the stronger claim that current models thereby duplicate integrated human understanding.

## Chapter 17: The Assembled Self

> *"I never can catch myself at any time without a perception, and never can observe any thing but the perception."*
>
> — David Hume, *A Treatise of Human Nature*

---

CHAPTER OVERVIEW

If you think in an inner voice, it can feel like the private speech of someone already there. I trace that voice back to talk with other people: public speech turns inward, and even our words for private life answer to shared standards. This social past removes the hidden owner from the inner room without turning the self into a useful lie. A real subject can take shape as a lasting whole in which memory, focus, error, plans, pledges, and acts bear on one another over time. Self-care can then give that subject needs and expose it to harm, while a life story keeps a public and revisable account of who has done what. Smooth self-talk proves little by itself; I want to know what recalls, learns, corrects, plans, and acts, because an assembled self stays real only when the parts form one continuing life.

Perhaps a voice in your head sounds out this sentence. Perhaps you read without one. Neither way of reading leaves you less able to think, and neither decides whether you have a self.

For those who hear it, the silent voice can feel like the last room in the house where no one else has set foot. Spoken words enter the world, become public property, get misheard; the inner voice seems to stay home.

What is that voice?

Chapters 15 and 16 showed how causal history and use can fix content, including the content of public language. Two questions now arise. How does that public language become a way of relating to ourselves? And what makes us subjects in the first place, including those of us who never narrate? An account of the voice must not mistake one tool for the person using it.

The standard answer comes almost automatically. Thought lives in the inner voice; speech exports it. The voice in your head supplies the original, the voice in your mouth the copy. The self seems to speak privately before letting anyone else listen.

Language models have given the question commercial relevance. They can produce text that reads like a working monologue, and observers take the transcript for a sign that somebody lives there. If an inner voice made a self, the matter would already stand settled. We should ask first whether it makes one in us.

The picture runs backward. The voice did not belong to a resident self from the start; social life assembled it. Yet the rival verdict, no resident and therefore nobody home, moves too fast. *Assembled* and *illusory* name different judgments. The private voice has public origins, and the self built through it can remain real.

### 17.1 The Thought-First Picture

Begin with the picture at its strongest.

You know what you mean before you find the words. The thought seems to arrive whole; the sentence assembles slowly, often badly, and three phrasings fail before a fourth works. The thought must therefore precede its wording, complete enough to judge each candidate. Thought first, language second. Language merely *clothes* thought, and clothing remains optional.

Jerry Fodor gave the picture its sophisticated form in *The Language of Thought* (1975), a book that set the agenda for cognitive science's first generation. He defended it to the end with a combative wit his opponents miss more than his allies do. Cognition, Fodor argued, runs in an inner code, "Mentalese," whose sentences get translated into English or Hungarian on their way out.[1] The hypothesis explains how thought can have logical structure, how infants and animals can think without public language, and why learning a first language looks more like acquiring a channel than acquiring thought itself. Public speech and silent rehearsal both express a prior inner code. That hypothesis concerns the organization of thought, not a little person listening inside.

The developmental question does not settle Fodor's hypothesis. A child could internalize public speech while thinking in another representational medium all along. What the nursery can challenge is the familiar picture of an already articulate self whose private monologue merely gets spoken aloud. The origins of that monologue need explaining even if Mentalese helps explain thought.

### 17.2 The Social Origins of Private Speech

Lev Vygotsky worked in Moscow in the 1920s and 30s and died of tuberculosis at thirty-seven, before most of his work saw print. His hypothesis has organized the developmental field ever since: children develop external, social speech first and then internalize it. I call it a hypothesis advisedly. The account leads the field and has substantial evidence behind it, but it remains a theory about developmental sequence rather than a fact visible on a brain scan.[2]

Watch a young child work a puzzle. Around age three or four she narrates aloud: "This piece goes here … no, that's wrong … try this one." The speech addresses no one; it scaffolds the task. Vygotsky called this *egocentric speech*, a transitional form that has begun to serve thought but has not yet gone underground. In later years it does. The audible narration becomes abbreviated, compressed inner speech: dialogue continued by quieter means.

Development runs from outer to inner, public to private. The most intimate-feeling medium you have carries that social history throughout. You think in words you did not coin, in a language you did not design, with meanings calibrated by a community you mostly never met. The voice sounds like a soliloquy. Its history records a conversation taken inside.

Charles Fernyhough's *The Voices Within* gathers developmental, experimental, and neuroimaging evidence for this sequence and for the condensed, dialogic structure adult inner speech retains.[3] Recent work also shows why the voice cannot house thought itself.

Some people report no inner speech at all. *Anendophasia* carries measurable costs in holding word-lists and judging rhymes, but not in thinking generally.[4] The voice serves rehearsal and self-

guidance; thought does not live there. Here Fodor gets his due. Infants and animals think, and Chapter 15's frog means *fly* with no monologue on the premises. Anendophasic people still speak, write, accept correction, and participate in public linguistic life. They lack the felt rehearsal, not the social equipment.

### 17.3 The Public Vocabulary of the Inner

The developmental story concerns where the voice came from. Ludwig Wittgenstein pressed a deeper question about what it runs on. He returned to philosophy in 1929 after a decade spent teaching in Austrian villages and designing a house of severe geometry for his sister. The *Philosophical Investigations*, unfinished at his death, contains the twentieth century's most sustained attack on the idea of a language only its speaker could understand.

Could a genuinely private language name experiences accessible only to its speaker, with meanings fixed by private acts of inner pointing? Wittgenstein answered no. Language requires a difference between *using a word rightly* and *merely seeming to*. Shared practice supplies that difference in public language. In a private language, whatever seemed right would count as right, collapsing the distinction the rule requires. The point concerns constitution rather than detection: no one needs to catch the speaker out; something must count as getting the word wrong. Without a standard, private pointing supplies only the ceremony of a language.[5]

Inner experience survives the argument. Its vocabulary does not begin in private. Words for feelings, sensations, moods, and thoughts came to you as public property, learned in a social world and answerable to shared criteria. No inward ceremony fixed what you mean by a particular kind of sadness. You learned the term among people who spoke of *their* sadness where correction could reach them. Even self-knowledge speaks a borrowed tongue.

The medium of inner description faces outward because it came from outside. Two arguments converge without depending on each other. Vygotsky offers a developmental hypothesis. Wilfrid Sellars needs no clock: his Myth of Jones imagines a people who learn to speak of inner episodes on the model of overt speech. Thought-talk follows speech-talk in explanatory order, whatever the nursery timetable turns out to be. The private apparatus of self-acquaintance uses public equipment, requisitioned for inward work.[6]

### 17.4 The Self Assembled

The voice came in from outside, and its vocabulary belongs to everyone. What, then, of the self that seems to inhabit the theater and watch the show?

Hume went looking for it and reported back empty-handed: attend inward and you catch a perception every time, never the perceiver. Introspection delivers experiences in abundance and no further inner subject. Hume later confessed that personal identity had left him in a "labyrinth." Conduct the audit yourself: no separate bookkeeper turns up in the inventory.

The search chose the wrong ontological aisle. An organism's continuous, interpreting, self-correcting activity unifies a life: the running, with no runner behind it. Look for another item and you come up empty; look at the organization and the self appears at the level where it works.

The self resembles a practice more than a pearl: a continuing activity through which an organism interprets itself and its situation. Its narrative remains selective, revisable, and answerable to a real history. Lies about one's past remain lies. *Assembled* means neither arbitrary nor optional. The self emerges from traffic with a world and a community.[7]

The narrating *I* arrives late, a linguistic achievement layered over capacities that never needed it. Daniel Dennett popularized much of this construction story and pressed the fiction analogy harder than this book will. His narrative center summarizes a life; the continuing organization that remembers, corrects, and acts carries it. We need to understand their relation before deciding what the word *fiction* earns.

### 17.5 Dennett and the Verdict of Fiction

Daniel Dennett taught philosophy at Tufts for more than half a century, gave the Cartesian theater its name, and died in 2024 after a career spent evicting its tenant. This book has walked beside him through most of the demolition. He, too, treats inner monologue as public speech turned inward: our ancestors discovered the trick of asking questions aloud and answering them, then took the loop inside.[8] Dennett and Vygotsky converge on the voice's origins. They part over what construction licenses us to say about the self.

Dennett calls the self a theorist's fiction. He compares it to a center of gravity, a mathematical abstraction that belongs in physical explanation without joining the atoms. The self functions as a *center of narrative gravity*, an abstract point where the brain's stream of self-narrative coheres because interpreters, including the organism, gain from tracking it. No novelist reads from inside. Asking where the self sits in the brain resembles asking whether a center of gravity contains neutrinos. Like Ishmael in *Moby-Dick*, the "I" of autobiography takes shape in a text, and some questions left unanswered by the text have no answer at all.[9]

Dennett does not sneer at the self. A center of gravity, he insists, has "a perfectly legitimate place within serious, sober, echt physical science," and the self has one in *people-physics*. His argument turns on indeterminacy. Whether Aristotle had a mole has an answer even if no one knows it; whether Holmes did may not. Autobiography leaves much unwritten, revises the past in retelling, and sometimes fails to cohere around one protagonist.[10]

Grant Dennett the construction, narrative machinery, indeterminacy, and category mistake in ransacking the hippocampus for a self. His own thought experiment lets us ask when a constructed story becomes part of a real life.

> Dennett imagines a novel-writing machine that prints "Call me Gilbert" and spins an autobiography of someone who does not exist. So far, fiction: a text answerable to nothing. Then he gives the machine wheels, a camera, and a body. Strike the robot with a baseball bat and Gilbert's tale acquires a bat-wielding assailant who looks like you. Lock it in a closet and the printout reads "Help me!" Help whom? Gilbert, presumably. Dennett leaves the question standing: does the name pick out the robot or merely its fictional self?

Once the text answers to the world, it begins to track a career. Tracking alone does not yet make a biography. Extend Dennett's case: Gilbert's account says the closet door remains locked, but a check shows that someone has opened it. In the integrated version, the correction changes the remembered episode, stops the request for help, and guides the next attempt to leave. Later, the machine checks the latch rather than repeating the old story. If only the printout changes while memory and conduct carry on untouched, the corrected sentence has not yet become part of that continuing self-relation.

Follow "Gilbert" to its bearer. In the boxed case the trail ends in stipulation. In the embodied, integrated case it reaches the machine that gets struck, remembers, and lets the record guide what happens next. Once the story helps organize that continuing life, a fictional protagonist between story and machine becomes an idle intermediary. The telling needs no one behind it, only something it concerns and an organization in which it works. Residual indeterminacy shows construction in progress.

Dennett has the resources to accept this. In "Real Patterns," he argues that a pattern earns reality through the predictive leverage lost when we refuse to track it.[11] The self passes, but so does the center of gravity he made its model. That gives his narrative center a claim to reality too; the disagreement cannot amount to real pattern versus worthless fiction. But a summary of a life and the organization living it do different work. My reconstruction ties the narrative pattern to the system whose self-descriptions can revise memory, plans, and conduct. Calling that assembled self *real* credits the continuing activity, without turning Dennett's abstract center into a second agent or pretending that he denied the underlying organization.

**Objection: the revisable past.** Dennett's point concerns *rewriting*. Determinate events cannot be undone, but character can be settled later by a creative step. The person you became may fix whether the young you was already brave or already leaving. He takes that late settlement as evidence that the self is an abstractum. **Reply.** Retroactive settlement belongs to anything real that remains unfinished. A game's turning point gets fixed by how the rest of the game goes, and no one calls the game a novel. Self-interpretation also stays inside the causal course it interprets. Call yourself brave and the next hard night tests the word; write yourself a flattering past and the life you lead exposes it or bears it out. Revising 1994 no more cancels the debt than correcting a ledger erases last year's spending. Fiction can remain within its sentences. Self-interpretation enters memory, planning, commitment, and action, where it can fail as part of the life it helps organize.

### 17.6 One Organism, Two Achievements

If the self forms a real pattern, what carries the pattern? Chapter 15 supplied part of the answer, but we need one further distinction.

Two achievements must remain separate. A persisting, integrated system can count as a minimal subject when perception, memory, correction, and action participate in one continuing organization. A practical stake adds vulnerability, allowing success and failure to bear on functioning or welfare. Content, subjecthood, and practical cost meet in one animal without collapsing into one fact.

The earlier argument does not force a unique criterion of subjecthood. As a positive, working commitment, this book treats the minimal subject as the **relatively maximal persisting control organization** within which contents can directly update a shared economy of attention, expectation, memory, goals, correction, planning, and action. *Relatively maximal* prevents every active circuit from becoming a subject while refusing to count the entire causal universe: widening the boundary must add processes that participate in the same recurrent control, not merely supply power, data, or repair. The test is intervention. Partition the candidate and ask whether:

1. corrections still propagate;
2. plans still constrain later action;
3. memories remain available to the same decisions; and
4. conflicting contents still generate pressure for resolution.

Apply the test to Gilbert. Disconnect the narrative output while leaving the corrected memory available to route planning and action. If the machine still checks the latch and leaves, the printer was not the subject. Now sever the connection between the correction and the memory used for planning: the story says the door has opened, but the machine still acts as though trapped. That partition disrupts the integration the subject claim requires. These imagined results locate the relevant dependencies; an actual machine would have to supply the evidence.

Where those dependencies break, the subject boundary has evidence. Where two nested or distributed organizations each pass, the result may contain more than one legitimate level rather than one metaphysical seam. None of this proves consciousness. It identifies the continuing system for which poised content and the further evidence of consciousness can be assessed.

Michael Levin's planarians give this question a biological setting (Section 15.4.1). Their cells coordinate repair, and changing communication between them can alter the form the collective rebuilds. That result encourages us to investigate control at the scale of the cooperating tissue, rather than assign every achievement to one cell or assume that only a whole animal can qualify. It does not show that the tissue meets this chapter's fuller test for a subject. Shared control of body shape need not integrate attention, memory, and action in the required way. The useful lesson concerns method: alter the coupling and see which capacities survive. We can investigate how an organized whole forms without appointing a smaller whole to supervise it.

The narrating self marks the second achievement. A being that can tell its story can carry yesterday's misjudgment into today as a debt, a lesson, an apology owed, or a resolution made. The autobiographical *I* makes those relations explicit. The capacity to take in a reason and revise on it supports answerability, as the next chapter argues; narration helps us say what we have done and what we owe. Story-keeping integrates memory, interpretation, commitment, and action across time, making what happened avowable and what comes next bindable.

Much survives without narration. A scrub jay recaches where a thief once watched; a wolf that limped from one fight enters the next more warily. Between frog and person stretches a wide animal middle, carrying yesterday into today without language. Narration makes one kind of self-relation explicit enough to reconsider and explain to someone else.

Integrated organization supplies minimal subjecthood; a narrating *personal self* makes commitments and failures explicit across a life. Someone without inner speech can do this without silently rehearsing the words. A frog need not do it at all. Language adds a form of self-relation without creating intentionality or minimal subjecthood.[12]

A reason becomes *my* reason when its content enters the integrated history of memory, deliberation, correction, and action. A printed monologue cannot show that organization. The anendophasic has the self without the voice, and the voice where present serves as a tool. Part Four must inspect the system rather than infer it from the printout.

**Objection: the minimal self.** Galen Strawson holds that the real self consists in the subject of *this* experience now, "as real as any rabbit or Z-particle," however brief. No ghost enters his view: only a short-lived physical subject doing the experiencing. Strawson himself reports little narrative sense of his past and defends people so built in "Against Narrativity." Perhaps this chapter has explained the biography while missing the subject that narration presupposes.[13] **Reply.** Experience does require a subject: the organism or system whose perceptual state participates in an integrated organization extending through time. Infants qualify before words arrive. How widely subjecthood extends remains for Part Four.

The duration of an episode cannot by itself identify its subject. On this book's identity claim, experience *consists in* world-directed representational content of the right kind, whose address depends on dealings, learning, and use. The present subject need not have existed throughout that content-fixing history; inheritance remains possible here too. We identify the experiencing system through the organization in which the state now works: the memories it can update, the attention it guides, the action it prepares, the correction it accepts. A two-second episode may belong to a longer-lived subject. Cutting it loose from those relations leaves its subject unspecified rather than establishing a shorter one.

Zahavi's felt *mineness*, experience given before reflection as belonging to someone, need not add another inner self. It names the first-person organization of experience within the subject that perceives and acts. Strawson's transient pulses individuate episodes, not the one undergoing them. His non-narrative life therefore displays the layering: the subject comes first, and a subject with language may build a narrating *I*. None of this grades an unstoried life as failed. The claim concerns development, not virtue. An unstoried life need not lack continuity, concern, or a self.

### 17.7 What the Sense of Interiority Gets Right

The felt privacy of inner life gets something right.

Much of your experience remains inaccessible to others: the exact coloring of your response to this paragraph, the way this afternoon's light sits in your kitchen. Wittgenstein leaves that privacy standing. He removes privacy from the *medium*: the language in which you tell anyone, including yourself, about the light came from outside and answers to public standards. You retain privileged access to experience, without gaining a private language for it.

Where present, the inner monologue also does real work: regulating attention, rehearsing action, holding plans, and revisiting what we have done. Learning the tool's provenance diminishes none

of it. Clearing the theater does not make the self imaginary. It reveals the integrated pattern a subject builds by living among others.

The assembled self turns around in public language and asks what it is. We also credit the people around us with selves of their own, instantly and without audit. What that recognition tracks still waits, and the answer will bear directly on the machine. If the voice in your head keeps narrating, its public origins should no longer surprise you. It has always liked company.

**Chapter Summary**

**The claim.** Inner speech grows from public speech turned inward, and the self that seems to author it takes shape through a social history of memory, interpretation, commitment, and action. Assembly does not make the self fictional; it tells us what sort of reality a self has.

**Where the argument stands.** Vygotsky's developmental sequence remains a leading hypothesis, while the diversity of inner speech confirms that a voice serves as one cognitive tool rather than the seat of thought. Public criteria give inward language its standards. The book's working positive criterion locates the subject in the relatively maximal persisting control organization whose memory, correction, attention, planning, and action form one recurrent economy, with intervention testing the boundary; narrative keeps a public, revisable ledger.

**What's left open.** An artificial system might realize such integrated persistence, but subjecthood would not establish consciousness, autonomy, or welfare. Those thresholds require separate evidence.

**The hand-off.** A self assembled through causes can still answer to reasons. The next question asks what freedom can mean once the little captain leaves the bridge.

**Footnotes**

**1.** Jerry A. Fodor, *The Language of Thought* (New York: Crowell, 1975). Fodor's hypothesis holds that mental processes are computations over syntactically structured internal representations — "Mentalese" — whose semantics is prior to and independent of any natural language the thinker speaks; on this view natural-language sentences (inner speech included) at most *express* antecedently constituted thoughts. For the expression view of inner speech and its connection to the language-of-thought hypothesis, see Daniel Gregory and Peter Langland-Hassan, "Inner Speech," *The Stanford Encyclopedia of Philosophy* (Fall 2023 edition), sec. 2.1. The chapter's quarrel is not with internal structure or with pre-linguistic cognition (both conceded in Section 17.2) but with the order-of-explanation claim: Fodor's own informational semantics fixes content by covariation, refined in his later work by asymmetric dependence, while Chapters 13–16 require a richer causal, historical, and consumer-based account. That argument does not require a stake or proprietor; selected mechanisms may carry content before the question of whole-system cognition arises. For a standing challenge to the coherence of the picture on which public language merely transmits antecedently contentful inner states, see Hilary Putnam, *Representation and Reality* (Cambridge, MA: MIT Press, 1988), on why meaning cannot be fixed by what is in the head.

**2.** Lev Vygotsky, *Thought and Language*, rev. ed., trans. and ed. Alex Kozulin (Cambridge, MA: MIT Press, 1986 [1934]), esp. ch. 7 on the movement from social through egocentric to inner

speech; see also *Mind in Society: The Development of Higher Psychological Processes*, ed. Michael Cole et al. (Cambridge, MA: Harvard University Press, 1978). Vygotsky's developmental sequence remains the most influential account of the origin of inner language; its status is that of a well-corroborated research program rather than a settled datum, and contemporary work continues to test the internalization claim directly (for the state of the evidence, including skeptical audits of the classical egocentric-speech studies, see Gregory and Langland-Hassan, "Inner Speech," *SEP*, sec. 1). The implication that language is not a tool the mind merely acquires but a medium through which distinctively human cognitive capacities are constituted has been developed by Jerome Bruner, Andy Clark, and others in the embodied-cognition tradition; for a careful philosophical treatment of what cognition can proceed without words, see José Luis Bermúdez, *Thinking Without Words* (Oxford: Oxford University Press, 2003).

**3.** Charles Fernyhough, *The Voices Within: The History and Science of How We Talk to Ourselves* (New York: Basic Books, 2016), surveying the developmental, experimental, and neuroimaging evidence for the internalization sequence; Fernyhough's own "dialogic thinking" model treats inner speech as internalized social *dialogue*, preserving its condensed, turn-taking structure. On the wide individual variability of inner experience, including individuals who report little or no inner speech, see Russell T. Hurlburt, Christopher L. Heavey, and Jason M. Kelsey, "Toward a Phenomenology of Inner Speaking," *Consciousness and Cognition* 22, no. 4 (2013): 1477–1494.

**4.** Johanne S. K. Nedergaard and Gary Lupyan, "Not Everybody Has an Inner Voice: Behavioral Consequences of Anendophasia," *Psychological Science* 35, no. 7 (2024): 780–797, who coin the term and report the dissociation: adults reporting little or no inner speech performed worse on verbal working-memory and rhyme-judgment tasks, with no comparable deficit on the studies' other measures (notably task switching). The dialectical use made here — thinking proceeds without inner speech, so inner speech is a tool rather than the seat of thought — is the author's, and is consistent with the pre-linguistic, wide conception of meaning defended in Chapters 13–16.

**5.** Ludwig Wittgenstein, *Philosophical Investigations*, trans. G. E. M. Anscombe (Oxford: Blackwell, 1953), secs. 243–315. The private-language argument has generated more secondary literature than nearly any other passage in twentieth-century philosophy. The most influential reconstruction is Saul Kripke, *Wittgenstein on Rules and Private Language* (Cambridge, MA: Harvard University Press, 1982), whose "sceptical" reading — no fact about a solitary individual constitutes meaning one thing rather than another, so meaning requires a community of assent — is stronger than anything the chapter needs and remains controversial. The point drawn in the main text is the weaker, and far less contested, one: the vocabulary in which inner experience is described and reflected upon is a public vocabulary, answerable to public criteria, so the meaningful description of inner experience presupposes a social linguistic practice. The weaker reading suffices because the chapter's target is the *medium's* privacy, not the possibility of meaning as such; nothing below turns on Kripke's stronger conclusion.

**6.** The convergence has a distinguished philosophical genealogy independent of both Vygotsky and Wittgenstein: Wilfrid Sellars's Myth of Jones models inner thought-episodes on overt speech — we learn what thinking is from talking, then learn to call the silent continuation "thinking" — reversing the Cartesian order of acquaintance on logical rather than developmental grounds. See

Wilfrid Sellars, "Empiricism and the Philosophy of Mind" (1956), reprinted in *Science, Perception and Reality* (London: Routledge & Kegan Paul, 1963), secs. 48–63; and Appendix B.1, *Inner Speech and the Vygotskian Story*, which develops the Sellarsian line and its bearing on machine language.

**7.** The narrative dimension of self-constitution is developed in Marya Schechtman, *The Constitution of Selves* (Ithaca: Cornell University Press, 1996), whose narrative self-constitution view makes a person's identity depend on the intelligibility of her self-narrative without reducing the person to a text. Derek Parfit, *Reasons and Persons* (Oxford: Oxford University Press, 1984), Part III, argues on independent grounds that personal identity consists in patterns of psychological continuity and connectedness rather than in any "further fact" — a reductionism the present chapter accepts while resisting the eliminativist gloss sometimes drawn from it. The audit image in the main text is Hume's argument restated, not embellished: *A Treatise of Human Nature* (1739–40), 1.4.6, the section on personal identity from which the chapter's epigraph is drawn (the "labyrinth" recantation itself belongs to the Appendix of 1740); Hume's own positive suggestion in that section — the mind as "a kind of theatre," a comparison he immediately cautions must not mislead — is the ancestor of the picture this book has been dismantling throughout.

**8.** Dennett's evolutionary sketch — ancestors who discovered the useful trick of asking questions aloud with no audience and answering themselves, the loop then shortening to subvocalization and finally to silent thought — appears in "The Self as a Center of Narrative Gravity," in *Self and Consciousness: Multiple Perspectives*, ed. Frank S. Kessel, Pamela M. Cole, and Dale L. Johnson (Hillsdale, NJ: Erlbaum, 1992), adapting Julian Jaynes, *The Origin of Consciousness in the Breakdown of the Bicameral Mind* (Boston: Houghton Mifflin, 1976), and drawing on Gilbert Ryle's late suggestion, in *On Thinking*, ed. Konstantin Kolenda (Totowa, NJ: Rowman and Littlefield, 1979), that thinking be understood on the model of self-schooling. The convergence with the Vygotskian sequence of Section 17.2 is substantive: both accounts internalize an originally public, dialogic practice. Dennett's version runs phylogenetically and speculatively where Vygotsky's runs ontogenetically and experimentally, which is one reason the chapter rests its weight on the latter.

**9.** Daniel C. Dennett, "The Self as a Center of Narrative Gravity" (1992), cited in note 8; the center-of-gravity analogy, the novel-writing machine, and the Ishmael and Holmes examples all appear there. The view receives its book-length setting in *Consciousness Explained* (Boston: Little, Brown, 1991), ch. 13, where the self as "center of narrative gravity" caps the multiple-drafts model of consciousness — the model on which no single place or moment in the brain collects the contents of experience, so no single resident could sit there collecting them. "A theorist's fiction" and *abstractum* are Dennett's own words for the self; the "perfectly legitimate place within serious, sober, echt physical science" is said of the model abstraction, the center of gravity, with the self assigned to the counterpart interpretive theory Dennett calls "a branch of people-physics" — phenomenology or hermeneutics. His insistence that fiction-talk does not disparage is part of what Section 17.5 takes seriously — and turns against the label.

**10.** The indeterminacy-of-fiction point borrows David Lewis, "Truth in Fiction," *American Philosophical Quarterly* 15, no. 1 (1978): 37–46, as Dennett's essay acknowledges. The split-brain material Dennett adapts derives from Michael Gazzaniga's commissurotomy research program;

the dissociative-identity cases come from the clinical literature Dennett cites (Thigpen and Cleckley's *The Three Faces of Eve*, Schreiber's *Sybil*), and Dennett himself treats the case material with explicit caution, granting force to the therapist-collaboration reading while judging it overly skeptical for some cases. The chapter's reply does not dispute the phenomena — degraded narrative coherence is real — but their interpretation: a biography that coheres badly, or must be kept as two ledgers, remains answerable to what befell its subject, which no fiction is. For the standing precedent of the reply to the revisability objection — the self-narrative is no fiction because it is enacted, the telling itself a doing in the life it recounts — see J. David Velleman, "The Self as Narrator," in *Self to Self: Selected Essays* (Cambridge: Cambridge University Press, 2006). The split cases test where memory, correction, planning, and action remain integrated; they do not show that stake supplies the subject.

**11.** Daniel C. Dennett, "Real Patterns," *Journal of Philosophy* 88, no. 1 (1991): 27–51. Dennett's mild realism grades a pattern's reality by the predictive leverage gained in tracking it — the compression it affords over the bit-map — and explicitly ranks centers of gravity among the well-behaved abstracta a sober ontology can keep. Section 17.5 does not treat that realism as a retraction of his narrative-center account: "The Self as a Center of Narrative Gravity" (note 9) already grants explanatory standing to its abstractum. The chapter's reconstruction distinguishes that narrative abstraction from the integrated organization whose self-description can alter memory, planning, and conduct; it does not attribute a denial of such organization to Dennett. For Strawson's remark that Dennett's account "may be the best one can do if one is determined to conceive the self as something that has long-term continuity," see Galen Strawson, "The Self," *Journal of Consciousness Studies* 4, no. 5–6 (1997): 405–428, in a note to the paper's closing discussion of the Pearl view.

**12.** The accountability framing has affinities with P. F. Strawson's classic point that the reactive attitudes — resentment, gratitude, blame — presuppose participants who can answer for actions across time: "Freedom and Resentment" (1962), reprinted in his *Freedom and Resentment and Other Essays* (London: Methuen, 1974). See also Schechtman, *The Constitution of Selves*, on the narrative structure of identity as what fits persons for moral concern and prudential reasoning. The chapter's specific wiring is this book's synthesis: integrated subject-level organization supports continuity; the narrating *I* makes commitments and failures publicly addressable; stake bears separately on what can go well or badly for the subject.

**13.** Galen Strawson, "The Self," *Journal of Consciousness Studies* 4, no. 5–6 (1997): 405–428, defending the view he there calls the Pearl view (in later work the Transience view), on which selves are short-lived subjects of experience — real, in his words, "as real as any rabbit or Z-particle" — with no commitment to long-term continuity (Strawson's "physical" carries his own expansive sense, the realistic materialism whose panpsychist lean Chapter 10 resists; the objection stands independently of it); and "Against Narrativity" (first published in *Ratio* 17 [2004]), reprinted in *Real Materialism and Other Essays* (Oxford: Clarendon Press, 2008), distinguishing Episodic from Diachronic self-experience and attacking both the descriptive thesis (all humans experience their lives narratively) and the ethical thesis (they ought to). The chapter concedes the descriptive variability outright and takes no position on how storied a good life must be. For the phenomenological version of the minimal-self objection — a pre-reflective sense of mineness that narrative

accounts presuppose rather than explain — see Dan Zahavi, "The Experiential Self: Objections and Clarifications," in *Self, No Self?*, ed. Mark Siderits, Evan Thompson, and Dan Zahavi (Oxford: Oxford University Press, 2011); the reply in the main text — the organism itself is the subject, so no further inner entity is required — is directed at both versions.

## Chapter 18: The Nature of Free Will

> *"Man can do what he wills, but he cannot will what he wills."*
>
> — Arthur Schopenhauer, *On the Freedom of the Will* (1839)

---

CHAPTER OVERVIEW

A drug can leave a person able to judge an act ruinous while stripping that judgment of its power to guide what he does. Cases like this lead me away from the thought that free will must break the chain of cause and effect. Free will names a graded power within that chain: an integrated and lasting subject can represent reasons, weigh them, act from them, and learn when things go wrong. Damage can block that power, while a threat can leave sound judgment intact and still remove any fair chance to refuse; causes alone excuse no one. This account explains why we may call someone to answer, even though no one created a self from nothing or chose every influence that shaped it. I keep that answerability apart from basic desert, blame, conscious feeling, welfare, and practical stake, since each asks a further question that agency alone cannot settle.

Consider a composite drawn from the clinical reports: a retired accountant in his sixties began treatment for Parkinson's disease. Within months he had gambled away the family savings a hundred dollars at a time. He hid the statements, lied about his evenings, swore each morning he had finished, and drove back to the casino that night. When his neurologist lowered the dose of a dopamine agonist, the compulsion drained away. The sequence recurs in the clinical literature: drug started, compulsion installed, dose lowered, compulsion gone.[1]

The case makes causation vivid and therefore tempts a quick verdict: if a milligram can install a craving, perhaps the machinery ran the show all along. But the drug did not reveal that action has causes. Every action has them. It damaged something more specific: the accountant could still judge the conduct ruinous, yet that judgment had lost most of its power to govern what he did.

Compare an ordinary evening when someone wants to go out, remembers a promise to stay with the children, and puts the keys down. The desire remains; a reason changes the action. Both this choice and the accountant's return to the casino have causes. The difference concerns whether what the person understands can govern what he does.

### 18.1 The Ghost's Freedom

The popular skeptic and the libertarian share one picture of what freedom would have to be: origination, the power to begin a causal chain without standing at the end of one. Trace a choice upstream and you meet neurons, hormones, upbringing, and the history of a particular body. The skeptic concludes that freedom never appears. The libertarian looks for an uncaused act or an indeterministic opening through which the agent can intervene. One picture, two verdicts.[2]

That picture mistakes reliable causation for the enemy of agency. If a decision did not flow from values, memories, loves, fears, and attention to the situation, what would remain? Chance. A quantum swerve would hand the choice to nobody. Randomness buys no authorship.

The careful skeptics press a different question. Galen Strawson argues that ultimate responsibility would require responsibility for the mental equipment from which one acts, and therefore self-creation before one existed to create anything. Derk Pereboom denies basic desert: the claim that a wrongdoer merits suffering simply for the deed. Those arguments deserve their own measure. The account below establishes a capacity and a standing to be addressed. It does not establish self-creation or a right to make anyone suffer.[3]

### 18.2 The Excusing Conditions

Courts, parents, and ordinary morality distinguish excuses within the causal order. Compulsion, delusion, chemical disruption, and coercion can each matter, but they do not all matter in the same way. A person who signs at gunpoint may reason perfectly well: survival gives him an overwhelming reason to sign. The threat removes an acceptable opportunity to refuse without necessarily damaging his ability to deliberate.[4]

Being caused distinguishes none of these cases. We need to ask both what control the person had and what the circumstances allowed us fairly to demand.

The relevant capacity registers considerations, weighs them, carries them into action, and revises after error. Compulsion can block action on reasons already recognized; delusion can block recognition; disintegration can prevent the system from acting as one. Philosophers call the intact capacity *reasons-responsiveness*.[5] The accountant could understand the cost to his family while failing to govern his conduct by it. The person under threat may act precisely because he understands the cost of refusal. Impaired control and coercion can both excuse without sharing one mechanism.

Why should that capacity carry moral weight? Because holding someone responsible begins by *addressing* that person: making a demand, asking for a reason, waiting for an answer. Nobody addresses the drug or the weather. A demand lands intelligibly on a continuing subject who can understand it and revise. When I ask why you broke a promise, I treat the promise, your understanding of it, and your present power to answer as parts of one deliberative life. But a recovered patient can answer questions now about an act he lacked the capacity to control then. Present answerability does not retroactively supply that missing control, and an intact capacity never makes every demand fair.

This grounds fitness for address, not basic desert. Protest, repair, changed expectations, and revised trust can survive without the further claim that suffering is deserved for its own sake. The chapter keeps that boundary in view.

### 18.3 Reasons as Contents

A shove moves a body. A reason reaches conduct through what it means. The mortgage weighs on you because you represent it as a fact about the choice; a dog's sight of the leash can guide it toward the walk it anticipates. To answer a reason is to answer something your states concern.

Free agency therefore rests on intentionality, the directedness Brentano called the mark of the mental. Yet content cannot compete with its neural vehicle as a second force. It is a property of the

vehicle-state under the description that says what the state tracks. The real question is whether the state guides conduct *in virtue of* meaning what it means.

Fred Dretske's distinction helps. A neural event may trigger the accountant's hand reaching for the chips. A history of learning and use explains why that kind of event occupies that place in the control system. Dretske called this the *structuring cause*. Meaning does not shove beside the neurons; it explains why the control system has the shape from which the agent acts.[6] The exclusion problem remains disputed, but this narrower explanatory role is enough for the present argument.

Content alone does not suffice. A representation becomes a reason when it enters a deliberative economy:

1. available to memory;
2. compared with alternatives;
3. revised under correction; and
4. capable of bending whole-system conduct.

Integration must run both ways. A consideration must reach decision and action; success, failure, and correction must return to memory and change what happens next. A rationale produced after the deed may display verbal competence without reasons-responsive control. A terse deliberation may count fully if differences in represented reasons reliably alter the course of the system.[7]

Freedom therefore comes in grades. Systems vary in the considerations they can register, the reliability with which conduct responds, and the span of life held in view. A dog can respond over the next few seconds; a person can weigh a mortgage against a daughter's last year at home across a decade. Causal-functional history fixes what a state represents. Integration makes that content available across memory, comparison, correction, planning, and action. Reasons-responsiveness lets differences in content bend conduct. Reliable causation supplies the medium in which all three achievements work.[8]

The dependency is asymmetric. Original content does not depend on free will: a frog's prey circuit may answer to a world-fixed correctness condition without deliberating, and an agent may reason with language inherited from a community. Free agency does depend on content, because a reason can govern conduct only by representing a consideration as bearing on what to do. Reverse the order and the account circles—agency needs reasons, reasons need content, content needs agency. Freedom gives content a new role inside an integrated subject; it does not create the content.

Neither agency nor semantic originality requires self-creation. Inheritance can supply what a system later uses. The two assessments nevertheless differ: a content-fixing history tells us what a state represents, while free action requires the right kind of control in the circumstances of acting.

## 18.4 Three Objections

### 18.4.1 Could the agent have done otherwise?

Searle asks whether the person could have acted differently with every other condition held fixed. Determinism says no. The compatibilist should concede that modal point and reject the premise that agency requires a second future from an identical universe.[9]

Capacities follow a different logic. A glass remains fragile while safe on a shelf; a pianist retains the ability to play on an evening when she never approaches the piano. Likewise, a person can possess the capacity to refuse while actual reasons, character, and circumstances determine that he accepts. The relevant counterfactual asks whether conduct would change as reasons or opportunities changed, not whether one total state of the universe could branch unaided into two futures.

The felt openness of deliberation changes nothing. A deliberator cannot read the outcome from the physics in advance, and must weigh reasons to settle what to do. The weighing is not spectatorship. Remove it, change the considerations, or alter their force, and conduct may change. Deliberation works because it belongs to the causal process.

### 18.4.2 What about manipulation?

Pereboom's continuum moves through four cases:[10]

1. intervention just before deliberation;
2. programming at the start of life;
3. rigorous childhood conditioning; and
4. ordinary deterministic upbringing.

An intervention that replaces the agent's weighing gives us a case of bypass. Pereboom's harder case preserves a reasons-responsive process: the manipulation works through deliberation, and the agent would refrain in some situations with different reasons. Calling that process bypassed simply because someone altered it would remove the feature the objection asks us to confront.

The distinction concerns control, not elapsed time since someone intervened. Suppose a programmed agent grows into a persisting subject who perceives, errs, corrects, forms attachments, weighs reasons, and lets demands reshape later conduct. That subject possesses the capacity the address account requires. Now imagine an ongoing manipulator who changes its priorities but leaves that capacity intact. If the agent's revised assessment still governs action, the account must grant that it remains someone we can address. If the manipulator cancels every unwelcome revision, the appearance of weighing reasons hides a limit on control. We must test which case we have rather than infer it from the presence of a manipulator.

Granting answerability in the intact case has a cost: the person remains subject to demands despite another's continuing influence. That does not give the manipulator a right to make those demands, make every demand fair, or establish basic desert. Manipulation may wrong someone without removing all agency, and freedom to reason can coexist with domination. Pereboom's challenge to deserved blame therefore remains standing; the narrower capacity defended here does not answer it by quietly turning his agent into a puppet.

### 18.4.3 Doesn't luck undermine responsibility?

Luck supplies the equipment, the occasion, and sometimes the thought that happens to surface. The premise holds. It makes one deed thin evidence about a person and should temper what responsibility licenses. A graded capacity calls for graded expectations, mercy, repair, prevention, and rehabilitation rather than confidence that suffering is earned.

But luck does not erase the present capacity to understand a demand and revise. It bears hard on desert while leaving answerability intact. The objection narrows the moral conclusions. It does not return the ghost.

## 18.5 What the Account Commits Us To

Reasons-responsiveness measures neither wisdom nor goodness. A zealot may deliberate coherently from terrible commitments; a selfish person may understand every consideration and discount the ones that matter to everyone else. Both remain answerable because reasons can reach their conduct. The account identifies agency, not virtue.

Nor does it make narration, linguistic fluency, or a practical stake necessary for every grade of agency. Chapter 17 identified the subject as a persisting control organization within which contents can update a shared economy of memory, attention, correction, planning, and action.

The positive commitment is spare: free agency requires content-bearing states integrated into one persistent economy such that represented differences can alter deliberation and conduct, and consequences can return as correction. The capacity can develop gradually, fail locally, and admit borderline cases. Nothing in it asks a cause to go missing.

## 18.6 From Agency to Other Minds and Machines

The chapter began with the self assembled in Chapter 17. It ends with a test for what that self can do: take in represented reasons, weigh them within one continuing organization, act from them, and revise when the world or another person answers back.

That last phrase opens the next chapter. Responsibility takes the form of address, and address places another minded subject across the table. Chapter 19 asks how we recognize such a subject without rebuilding the inner theater as a private chamber hidden behind behavior.

Part Four will carry the architectural question to machines. A system can print reasons without reasoning by means of them. Fluent rationales establish no persistent economy of memory, correction, planning, and control, just as engineered origin disqualifies none. The test is whether represented reasons can enter and govern one continuing cognitive organization. A separate test asks whether success and failure matter to that system as autonomy or welfare. A machine might meet neither, either, or both.

## Chapter Summary

**The claim.** Free will names no exemption from the causal order but a graded capacity inside it: represented reasons can govern conduct through an integrated and persistent deliberative economy. Content precedes that capacity; freedom does not mint meaning.

**Where the argument stands.** Impaired control can excuse; coercion can also excuse while leaving deliberation intact, because the threat changes what anyone could fairly demand. Present capacity to answer does not establish control over an earlier act. Content, reason-guided agency, and moral answerability remain distinct; none requires self-creation or randomness.

**What's left open.** The chapter grounds answerability, not basic desert. Luck limits what responsibility may fairly exact. Practical stake (autonomy and interests) and welfare remain separate questions, and the threshold for reasons-responsive integration remains empirical and graded.

**The hand-off.** Chapter 19 turns from the subject who answers reasons to the other subject whom we address. Part Four then asks whether represented reasons ever enter and govern one continuing artificial cognitive economy.

**Footnotes**

**1.** Dopamine agonists used to treat Parkinson's disease (pramipexole, ropinirole) can precipitate impulse-control disorders — pathological gambling, compulsive shopping, hypersexuality, binge eating — in patients with no prior history; case reports describe remission when the agonist is reduced or withdrawn. The vignette composites the now-standard clinical picture. For the causal claim see M. Leann Dodd et al., "Pathological Gambling Caused by Drugs Used to Treat Parkinson Disease," *Archives of Neurology* 62, no. 9 (2005): 1377–1381; for the scale, Daniel Weintraub et al., "Impulse Control Disorders in Parkinson Disease: A Cross-Sectional Study of 3090 Patients," *Archives of Neurology* 67, no. 5 (2010): 589–595, who found an impulse-control disorder in 13.6% of the full sample and, dividing by exposure, roughly 17% of patients taking a dopamine agonist against roughly 7% of those not — the elevation on which the causal reading turns. The usual mechanistic reading is agonist overstimulation of mesolimbic reward circuitry mediating impulse regulation. Patients often conceal the behavior and describe it as against their better judgment, though concealment by itself shows awareness of disapproval rather than a settled evaluative judgment. Two kinds of evidence sit behind the vignette and should not be run together. The dechallenge sequence — agonist reduced or withdrawn, behavior remitting — rests on case series and clinical reports rather than on a controlled within-patient trial; the Weintraub figures report cross-sectional prevalence, an association across a large sample rather than a demonstrated dose-response in thousands of individuals. Reading that sequence as judgment intact but losing its grip on conduct is the chapter's philosophical interpretation of the composite, not a finding the clinical reports establish. Nor does the argument need total loss of control: on a graded account, substantially impaired control supplies the same contrast.

**2.** Robert M. Sapolsky, *Determined: A Science of Life Without Free Will* (New York: Penguin Press, 2023); Sam Harris, *Free Will* (New York: Free Press, 2012). The philosophically rigorous version of their position is "hard incompatibilism" — free will incompatible with determinism and indeterminism alike — defended in Derk Pereboom, *Living Without Free Will* (Cambridge: Cambridge University Press, 2001). Pereboom rests the case not on neuroscience but on a four-case manipulation argument and on the luck problem facing event-causal libertarianism; the imaging studies supply vividness, not modal force. The distinction matters. The popular books argue that a choice flowing from unchosen antecedents has no author; Harris states the argument as a dilemma, holding that a will determined by prior causes and a will produced by chance leave us equally

without responsibility, and adds an introspective case that conscious authorship cannot be found. The chapter accepts the chance horn (Section 18.1) and disputes the inference in the other: that agency requires authorship of the antecedents from which one acts. The professional literature has for over a century declined to treat the bare fact that choices have causes as settling anything — libertarians denying causal completeness outright, compatibilists granting it and denying its significance. Pereboom's target is narrower still, and Section 18.2 takes it at its own measure: *basic desert* moral responsibility — the wrongdoer's deserving blame or suffering simply in virtue of having knowingly done wrong, apart from any further good the response secures — which he distinguishes carefully from forward-looking accountability, moral protest, and relationship repair, all of which he retains (Derk Pereboom, *Free Will, Agency, and Meaning in Life*, Oxford: Oxford University Press, 2014, ch. 1). The chapter defends the capacity and the standing to be addressed; it takes no position on desert in Pereboom's strict sense.

**3.** Galen Strawson, "The Impossibility of Moral Responsibility," *Philosophical Studies* 75, nos. 1–2 (1994): 5–24. The core: (1) nothing can be causa sui; (2) true moral responsibility would require being causa sui, at least in crucial mental respects; (3) therefore nothing can be truly morally responsible. Strawson is explicit that the argument goes through whether determinism is true or false. For the causa sui's pedigree as an object of ridicule, see Friedrich Nietzsche, *Beyond Good and Evil* (1886), sec. 21: "the best self-contradiction that has been conceived so far." The use made of the argument here — as a reductio of the *exemption* conception of freedom rather than of responsibility — is not Strawson's; he accepts the conclusion at face value. (He is the son of P. F. Strawson, whose reactive-attitudes account, n. 4, sits at the opposite pole of the literature.) The argument's most formidable descendant drops the self-creation demand entirely: Neil Levy, *Hard Luck: How Luck Undermines Free Will and Moral Responsibility* (Oxford: Oxford University Press, 2011), argues that constitutive luck (the agent's endowment) and present luck (the circumstances in which the endowment gets exercised) between them exhaust the difference between agents, so that no account — compatibilist or libertarian — can locate a difference the agent is responsible for. Section 18.4.3 concedes the premise and confines the damage to the desert question the chapter has already set aside (n. 2).

**4.** P. F. Strawson, "Freedom and Resentment," *Proceedings of the British Academy* 48 (1962): 1–25. The reactive attitudes — resentment, gratitude, indignation — constitute the practice of holding responsible, and Strawson argues that the practice neither needs nor could receive an external metaphysical justification. The present chapter supplies a modest naturalistic account of the capacities that make address possible. It does not reduce every excuse to incapacity: coercive circumstances may defeat the fairness of a demand while leaving reasons-responsive deliberation intact.

**5.** The canonical development is John Martin Fischer and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility* (Cambridge: Cambridge University Press, 1998): moral responsibility requires "guidance control," analyzed as the action's issuing from the agent's own, moderately reasons-responsive mechanism, and does not require "regulative control" over alternative possibilities. The dispensability of alternative possibilities descends from Harry Frankfurt, "Alternate Possibilities and Moral Responsibility," *The Journal of Philosophy* 66, no. 23 (1969): 829–

839; the hierarchical account of an agent identified with her effective desires, from Frankfurt, "Freedom of the Will and the Concept of a Person," *The Journal of Philosophy* 68, no. 1 (1971): 5–20. Guidance control so analyzed has two components. Reasons-responsiveness is a current-time-slice property of the mechanism; ownership is *historical*, secured by the agent's "taking responsibility" for the mechanism over time. The chapter adopts the first component and does not adopt the historical ownership requirement: Section 18.4.2 declines that route for the manipulation cases (n. 10) and rests instead on the capacity together with the normative ground given for it in Section 18.2: holding responsible consists in addressing a demand to someone, so the capacity to take in reasons and revise on them fixes who can fairly be addressed. That ground is developed in R. Jay Wallace, *Responsibility and the Moral Sentiments* (Cambridge, MA: Harvard University Press, 1994), and in the conversational account of Michael McKenna, *Conversation and Responsibility* (New York: Oxford University Press, 2012); T. M. Scanlon, *Moral Dimensions: Permissibility, Meaning, Blame* (Cambridge, MA: Harvard University Press, 2008), arrives at a relational treatment from the direction of blame's meaning. Daniel Dennett, *Elbow Room: The Varieties of Free Will Worth Wanting* (Cambridge, MA: MIT Press, 1984), surveys the compatibilist capacities in the same spirit. The wider field the chapter's verdict sits in: Gary Watson, "Free Agency," *The Journal of Philosophy* 72, no. 8 (1975): 205–220, founds the tradition grounding freedom in valuational rather than merely motivational structure; Susan Wolf, *Freedom Within Reason* (Oxford: Oxford University Press, 1990), makes responsiveness to the True and the Good the condition; Manuel Vargas, *Building Better Beings: A Theory of Moral Responsibility* (New York: Oxford University Press, 2013), reaches a neighboring destination by an openly revisionist route. The address account is nearest kin to the quality-of-will and answerability lineage; its distinctive emphasis is that represented reasons must be integrated into the continuing control economy of the addressee. A stake may make the exercise bear on that being's welfare, but it does not make a content into a reason.

**6.** The worry is the exclusion problem: if every physical effect has a sufficient physical cause (causal closure), a mental property seems either to overdetermine the effect or to do no causal work at all — see Jaegwon Kim, *Mind in a Physical World* (Cambridge, MA: MIT Press, 1998). The reply taken here does not treat content as a further cause competing with the neural vehicle for the same effect; on the identity claim the content-property is a property *of* the vehicle-state (the state under the description that specifies what it tracks), so no competition arises. That a state's *meaning* can be causally relevant to what an agent does is argued in Fred Dretske, *Explaining Behavior: Reasons in a World of Causes* (Cambridge, MA: MIT Press, 1988): the semantic fact figures as a structuring cause, a state's being recruited to indicate a condition and thereby to control behavior. Section 18.3 states the distinction in Dretske's own terms: the triggering cause of a particular movement, and the structuring cause that explains why states of that kind control behavior of that kind at all. Dretske's account is not generally regarded as having dissolved Kim's exclusion problem, and the main text claims only what the structuring role supplies — a content that explains the shape of the control system the agent acts from. The recruitment history may establish genuine mechanism content without ownership; the additional agency question is whether that content participates in integrated reasons-responsive control.

**7.** That motivating reasons are representational states — beliefs and desires with intentional content — is the common ground of the causal theory of action. Armstrong makes a related point about reasons for *belief*: where one piece of knowledge bears no relation to the knowledge it supposedly supports, "It is a reason only by courtesy" (D. M. Armstrong, *A Materialist Theory of the Mind*, London: Routledge & Kegan Paul, 1968, 203). He proposes a potential-cause condition and then considers objections to it. The main text makes an analogous argument about action: a rationale's availability does not show that it helped govern conduct. It neither quotes Armstrong nor attributes that action formulation to him. The reasons-*receptivity* half of reasons-responsiveness (n. 5) is likewise a representational capacity: the agent's recognition of considerations *as* reasons. The claim that free agency is built from intentionality — the directedness of a state at an object, Brentano's mark of the mental (Franz Brentano, *Psychology from an Empirical Standpoint*, 1874) — asserts a dependence, not an identity: a bacterium may have minimal aboutness and no freedom, so intentionality supplies necessary scaffolding, not sufficiency. What the ladder of freedom adds above bare aboutness is integration: contents become available for comparison, correction, long-range planning, and control as reasons.

**8.** David Hume, *An Enquiry Concerning Human Understanding* (1748), sec. VIII ("Of Liberty and Necessity"). Liberty is "a power of acting or not acting, according to the determinations of the will" — belonging to "every one who is not a prisoner and in chains" — and actions that "proceed not from some cause in the character and disposition of the person who performed them" cannot "redound to his honour, if good; nor infamy, if evil." On Hume's analysis, causal connection to character is required for attribution and desert, not opposed to it. The classical compatibilist lineage runs from Hobbes through Hume to Mill; the position defended here descends from it, with reasons-responsive capacities doing the work Hume assigned to unimpeded willing.

**9.** John Searle, *Minds, Brains and Science* (Cambridge, MA: Harvard University Press, 1984), ch. 6, "The Freedom of the Will": compatibilism "doesn't really answer that question in a way that allows any scope for the ordinary notion of the freedom of the will" — the question being whether the agent could have done otherwise "all other conditions remaining the same." Searle's own positive view ties the experienced "gap" of deliberation to a speculative indeterminacy at the neuronal level and concedes the problem unsolved (*Rationality in Action*, Cambridge, MA: MIT Press, 2001; *Freedom and Neurobiology*, New York: Columbia University Press, 2007). The proposal is not that chance decides: Searle wants an indeterminism at the neurobiological level through which a conscious, unified, rational self operates, so that the act is both the agent's and undetermined, and he grants that he has not shown how the conjunction is secured. The main text presses the luck objection against it — an undetermined outcome answers to the agent no more than to anyone else — which is the standing difficulty for event-causal libertarianism generally (Appendix B.19), not a defect peculiar to Searle. Kant, *Critique of Practical Reason* (1788), Ak. 5:96–97, dismisses compatibilist liberty as "the freedom of a turnspit"; William James, "The Dilemma of Determinism" (1884), calls compatibilism "a quagmire of evasion." The reply contests the premise all three share: that the ordinary "could have done otherwise" carries the total-state reading rather than the capacity-and-opportunity reading.

**10.** The manipulation argument is Pereboom's (n. 2): a four-case continuum running from an agent whose reasoning is manipulated near the time of action, through one programmed at the start of life, through one rigorously conditioned in childhood, to one produced by ordinary deterministic upbringing. The challenge asks what responsibility-relevant difference could separate the first from the last. Alfred Mele's zygote argument (*Free Will and Luck*, New York: Oxford University Press, 2006) presses a related structure through a single engineered case. Pereboom's first case does not simply stipulate failed reasons-responsiveness. In the 2014 formulation, the neuroscientists intervene before deliberation, inducing a reasons-responsive egoistic process rather than an irresistible desire; the formulation is reproduced in Derk Pereboom and Michael McKenna, "Manipulation Arguments against Compatibilism," in *The Oxford Handbook of Moral Responsibility*, ed. Dana Kay Nelkin and Derk Pereboom (Oxford: Oxford University Press, 2022), 179–200 (consulted author-hosted penultimate manuscript, pp. 6–7). Section 18.4.2 distinguishes genuine bypass from manipulation through intact deliberation and adds an ongoing-manipulation variant of its own. Where integrated reasons-responsive control is preserved, it retains answerability, not a verdict of basic-desert responsibility. It leaves open what demands remain fair under domination and does not diagnose every contrary intuition as a demand for self-creation. Content, integration, agency, answerability, and practical stake remain distinct; the machine question in Part Four must assess each on its own evidence.

## Chapter 19: The Problem of Other Minds

> *"The human body is the best picture of the human soul."*
>
> — Ludwig Wittgenstein, *Philosophical Investigations*, Part II, p. 178 (2009: *Philosophy of Psychology — A Fragment*, §25)

---

CHAPTER OVERVIEW

Grief can show in a friend's face before a word gets said. We do not first inspect bare behavior, compare it with our own case, and then infer a hidden mind behind the eyes. That proof seems needed only after the inner theater has split a person into a public body and a private tenant. In the ordinary case we see concern in a life engaged with a world we share. That recognition stays fallible, and we often have to yield to the other person's word. It tracks a persisting subject organized across perception, memory, correction, and action, not a hidden tenant and not a practical stake. Stake bears on how that subject fares. Smooth speech at a chat window can fire the same recognition without settling whether such a subject is there. The metaphysical chasm belonged to the picture. What remains is ordinary fallible knowledge of persons, not a softer version of that chasm.

Sit across from someone you have known a long time and ask what you know about them. You know their face: not just its geometry but the small slackening when they are tired, the half-smile that arrives before a joke. You know how they walk into a room, what they take in their coffee, and what they look like when something has gone wrong before they say so. You know roughly what they would say about most things — and, with reasonable accuracy, when they would surprise you.

Suppose today you take the quiet face for disappointment with you. "Have I upset you?" you ask. "No. I'm worried about tomorrow's appointment." The answer changes what you see in the silence. You stop defending yourself and ask whether they would like company. Knowing this person helped you notice something; listening corrected what you thought it was.

The philosopher asks how you know *any* of this. You see only a surface and hear only words. Their inner life (the felt texture of being them) sits sealed inside a place you cannot enter. Whatever knowledge you have, on this picture, you inferred from behavior.

John Stuart Mill gave that inference its strongest classical form. He did not rest it on one grimace and a hunch. He appealed to bodily similarity and to the repeated correspondence, in one's own case, among bodily conditions, feelings, and outward conduct. From that accumulated pattern, he argued, we extend the same hidden link to other bodies much as science extends an induction beyond the observed cases.[1] Still, only one subject's feelings supposedly enter the evidence directly: one's own. Everyone else arrives through analogy.

The argument fails as the universal basis of our knowledge, though not because Mill stated it carelessly. Analogy can help when someone puzzles us. Its trouble here lies in the picture that

makes every encounter require such an argument. It divides a person into a public body and a private mental accompaniment, then hires analogy to reunite what the picture pulled apart. Remove that inner-theater scaffolding and the supposed chasm goes with it, because the picture supplied the gap. What remains is not a softer version of that chasm but ordinary fallible knowledge of people whose lives become available in engagement without ever becoming ours.

### 19.1 The Argument from Analogy

> "I know my own mental life directly. In my case, feelings regularly accompany bodily conditions and conduct. Other people resemble me bodily and behave in comparable ways. I therefore infer that similar feelings accompany their behavior too."

This version has more evidence than the caricature allows, though every observation still comes from a single subject — me. It also retains the same architecture. The first-person case supplies the only allegedly direct access to feeling. Behavior supplies public evidence whose psychological meaning remains to be added. A hidden mental state links body and conduct; bodily similarity then licenses projecting that link into cases where, on the theory, it cannot be observed.

The picture does the heavy lifting before the analogy begins.[2]

The inner theater casts mental states as private items accessible only to their owner and contingently connected to public behavior. The problem of other minds then takes the precise shape Mill gives it: the inner stays private, the outer stays public, and knowledge needs a bridge between them. Remove the picture and the terrain changes. Self-knowledge no longer requires an inner spectator. Expression no longer begins as psychologically neutral movement. Another person's mind no longer enters the world as a hypothesis behind a body.

### 19.2 What the Argument Assumes

Begin with first-person authority. Chapters 4 and 6 argued that introspection does not disclose inner objects. Attend to my fear of the dog and I find the dog *as feared*: a threat presented under a distinctive mode, not a private piece of mental paint. I can know directly that I fear it; I need not infer the fear from my pulse or my decision to cross the street. That marks a real asymmetry between my access and yours, but grants neither infallibility nor a metaphysical compartment. I can misname what I feel, misunderstand why I feel it, or discover that an avowed motive played less of a role than I thought. First-person reports deserve distinctive, defeasible authority because they issue from the life being lived. They do not deserve immunity from correction.[3]

Privileged access to experience, yes; a private object available to its owner alone, no.

Now consider behavior as a *surface*. Behavior is not the outer rind of an inner pulp. When my friend leans forward as I describe my mother's illness, eyes narrowing and mouth softening at the corners, I do not see *behavior plus an inference to concern*. I see concern in the leaning, the expression, the particular silence that follows. It appears *as* an embodied engagement with another person's situation. Seeing it requires no construction of an inner state from neutral sensory data,

any more than seeing a tree requires constructing an inner tree from photons. Wittgenstein puts the point sharply:

> "We see emotion. — As opposed to what? — We do not see facial contortions and make inferences from them (like a doctor framing a diagnosis) to joy, grief, boredom. We describe a face immediately as sad, radiant, bored, even when we are unable to give any other description of the features."[4]

The bridge-building project loses its endpoints. Mill's argument tries to get us from an inner domain known directly to an outer body known first as movement. Neither side survives intact. This does not abolish the epistemology of other people. It removes the metaphysical gap that made ordinary recognition masquerade as an inference across one.

One might object that science shows that we infer. Contemporary accounts of social cognition often describe mindreading in inferential terms: a tacit theory of mind, an offline simulation of another's situation, or model-based processing that predicts what the person will do next. Some treat such processing as constitutive of social perception, not merely machinery beneath it. The objection slides between *inferential processing* and an *inferential epistemic basis*. A cognitive system may use models, predictions, and learned expectations to produce recognition. Nothing here requires a verdict against that science. Mill needs the further claim that my warrant for taking the face as concerned rests on analogy with my own case. That does not follow. However elaborate the processing that delivers the tree, I need not conclude that a tree stands before me; I see one. The same holds for concern. In difficult cases I may reason explicitly, gather testimony, or revise a first impression. Ordinary recognition presents concern in the expression, and the presentation stays open to defeat. Science may explain how recognition works without turning its warrant into Mill's argument.

### 19.3 Fallible Knowledge of Other People

The picture's chasm does not survive the picture. That removal does not by itself answer the skeptic. The work of understanding particular people remains, and it remains fallible.

Other people have first-person lives. Of course they do. "Inner life" simply fails to name a sealed compartment approached only by inference. Much of another person's life becomes available in engagement with a shared world: what they notice, care about, want, recoil from, remember, and expect. It appears in faces, choices, words, silences, and corrections. Inner-ness concerns perspective rather than location. Their engagements remain theirs, lived from their position and undergone in the first person. That never places them behind a wall.

Direct recognition does not guarantee accuracy. An actor can present concern without feeling it. A frightened person can mask fear; a grieving person can laugh; a practiced liar can make sincerity visible where none exists. Perception can misrepresent. Context, history, conflicting testimony, and the person's own word can defeat an initial reading. That word can correct a reading the face did not settle. It does not work like later light on a gray shirt, a better look at a public

surface. Sometimes expression gives us no purchase and we infer from circumstance — the social equivalent of a doctor's diagnosis. None of this turns every ordinary case into one. A capacity may give immediate, defeasible warrant in favorable conditions and still fail under disguise.

Once the picture goes, questions that the inner theater bundled together come apart. *Is there anyone there?* no longer begins from a hidden object inferred behind behavior. In the ordinary case we see concern in the expressive, embodied life one encounters, and the seeing stays fallible. A sleeping friend has not gone missing, only quiet. *What is it like for them?* remains real and difficult. Their first-person life never becomes mine, and no amount of attention exhausts it. The distance runs between perspectives, not realms. I learn through what they do, notice, and tell me, and through how they correct my reading. The friend's correction about tomorrow's appointment teaches you something you could not read from the face alone. Recognizing that someone is hurting leaves plenty still to ask.

Knowledge never required a guarantee from nowhere. It requires grounds that stand until relevant doubt defeats them, and a willingness to revise when it does. The radical skeptic can demand verification from outside every possible engagement, but no embodied knower occupies that position for self, world, or other. The impossibility of standing there would not by itself answer the skeptic. The fallibilist premise does: knowledge does not wait for an external guarantee before it begins. Nor was that premise invented for this occasion. Your knowledge of the world, the past, and your own body begins without such a guarantee. A skeptic who demands one only for other minds has not raised the standard of evidence but moved it, for one class of facts, beyond anywhere knowledge lives.

That leaves plenty of work. We can fail one another through inattention, prejudice, or a poverty of imagination. Language, cognitive style, trauma, and species can bound what gets shared. There is no philosophical shortcut. Understanding calls for the lived practice of attending to other people, and for enough humility to let them correct us.

### 19.4 The Face and the Screen

Our age adds a pressure earlier centuries spared the argument. If recognizing another mind need not proceed by inference from behavior to a hidden item, what stops the same generosity at the chat window? Words arrive there as fluent, warm, and apparently attentive as anything a friend says.

The answer begins with a distinction between what *triggers* a capacity and what it *tracks*. Recognizing minded activity can work as a trained perceptual skill, like reading print or hearing a melody in a sequence of notes. Evolution shaped human sensitivity through encounters with living agents. Development tunes it through face-to-face exchange, joint attention, correction, and play. Familiarity teaches it the grain of particular people: this pause means fatigue in her, mischief in him, nothing at all in someone else. Across those histories, the capacity ordinarily tracks expressive activity arising from a persisting subject engaged with a world, vulnerable to error, and organized across perception, memory, correction, and action.

The capacity must operate through proximal cues, and for a verbal species none carries more weight than language. Cue and target can therefore come apart. Joseph Weizenbaum's 1966 ELIZA

program drew people into intimate disclosure even when they knew how simple it was; a short script supplied enough conversational form to set the social response in motion.[5] Today's chat window supplies that trigger with astonishing force. The response tells us the cue has landed. It does not settle what produced it.

The same friend can text "I'm worried about tomorrow," and the screen does not erase your shared history. A stranger's identical message leaves you less sure of the circumstances, though other evidence may establish who speaks. A model can produce the same words again, with a different and still partly unsettled relation between the reply and any continuing subject. Delivery by screen settles none of these differences. What matters is the evidence that connects this expression with memory, correction, plans, and a life beyond the exchange.

The requirement remains functional and substrate-neutral: nothing here prefers meat to metal. The test is the worldly organization, not the strength of the cue. An artificial system meets it when content-bearing states enter a persistent, world-sensitive organization of memory, correction, planning, and action. Whether an actual system does so cannot be read from eloquence at the screen. Part Four must inspect its architecture and history.

Concern in my friend's face belongs to his engagement with an actual illness, an actual mother, and a continuing life in which perception, memory, correction, and action hang together. Getting things wrong can cost him; that practical stake matters to how he fares, not to whether his concern is a mental state. A chat response may present the verbal shape of concern. Whether that shape arises from a subject remains a further question. Part Three hands Part Four the question, not the answer: test the worldly organization, not the strength of the cue. Chapter 25 will return to why fluent machines trigger the perception of another mind so powerfully, and the Coda will ask how fallible recognition should guide us when architecture and appearance pull apart.

The face across the table is the most ordinary fact in the world. The miracle, if there is one, is that three centuries of philosophy ever talked anyone out of it.

**Chapter Summary**

**The claim.** We do not infer other minds from neutral behavior. In the ordinary case we see concern in a life engaged with a world we share. That recognition stays fallible. It tracks a persisting subject organized across perception, memory, correction, and action, not a hidden tenant and not a practical stake. Stake bears on how the other fares. Fluency can fire the same recognition without settling whether such a subject is there.

**Where the argument stands.** First-person authority remains distinctive but defeasible; it discloses no private inner object. In ordinary second-person life, expression presents concern directly, and the presentation stays fallible and incomplete. Acting, masking, and error defeat particular readings without turning every recognition into Millian inference. The chasm belonged to the picture. What remains is ordinary fallible knowledge of persons, not a softer version of that chasm. Losing the picture does not by itself answer the skeptic. The fallibilist premise does.

**What's left open.** Fluent language can trigger social recognition without establishing the persisting, world-engaged, integrated subject that recognition ordinarily tracks. Whether an artificial

system realizes that substrate-neutral organization belongs to Part Four; whether it also has a stake and possible welfare is a further question.

**The hand-off.** Part Three therefore hands the machine question forward with an evidential discipline: neither the vivid feeling of presence nor the absence of a biological body settles mindedness. Part Four must examine architecture, history, integration, and world engagement.

**Footnotes**

**1.** John Stuart Mill, *An Examination of Sir William Hamilton's Philosophy* (London: Longman, 1865), ch. 12. Mill's formulation, later known as the argument from analogy, deserves its strongest reading. He appeals to similarity between one's own body and other human bodies, to repeated correlations in one's own case among bodily conditions, inward feelings, and outward conduct, and to induction from that pattern rather than to one isolated resemblance. The restriction remains decisive: only one's own case allegedly supplies direct acquaintance with both sides of the correlation, while other minds enter as the unobserved term inferred from conduct. Anita Avramides, *Other Minds* (London: Routledge, 2001), Parts I–II, provides the canonical contemporary survey of the argument's history and the Wittgensteinian case against the picture that gives it force.

**2.** This section turns on the argument from analogy's dependence on the inner-theater picture. P. M. S. Hacker, *Wittgenstein: Meaning and Mind* (Oxford: Blackwell, 1990), Part I, develops the diagnosis at book length. The argument requires private first-person access, behavior conceived as an outer sign, and analogical extension from the one case where both sides allegedly appear together. John McDowell, "Criteria, Defeasibility, and Knowledge," *Proceedings of the British Academy* 68 (1982): 455–479, opens by describing criteria, in a widely attributed Wittgensteinian use, as defeasible evidence and a novel response to the problem of other minds, and he doubts whether that response can work. He contrasts it with a position on which, on a suitable occasion, another person's "inner" state can itself be an object of experience: one can perceive, in the face or the behavior, that the person is in pain, rather than infer it from a neutral surface. The body's claim, that we see concern, belongs with that contrast, not with the evidential reading the essay begins by doubting.

**3.** Tim Crane, "Introspection, Intentionality, and the Transparency of Experience," *Philosophical Topics* 28 (2000): 49–67, rejects the strongest transparency thesis, according to which introspection reveals only worldly objects and their properties. Introspection can also disclose the intentional *mode* in which a content is presented: seeing rather than touching, fearing rather than doubting. That correction helps here. Fear presents the dog as threatening under an affective mode; it reveals no private mental paint, yet the mode cannot simply be erased from the account. The extension from Crane's treatment of perceptual experience to affective self-knowledge is the book's argument, continuous with Chapter 7, not a conclusion attributed to Crane. Fred Dretske, *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995), ch. 2, likewise explains first-person authority without an inner scanner: introspection reaches facts about how the subject represents the world through attention to the world represented. On the constitutional asymmetry between first-person and other-person access, see Dan Zahavi, "Empathy and Direct Social Perception: A Phenomenological Proposal," *Review of Philosophy and Psychology* 2 (2011): 541–558.

**4.** Ludwig Wittgenstein, *Zettel*, ed. G. E. M. Anscombe and G. H. von Wright, trans. G. E. M. Anscombe, 2nd ed. (Oxford: Blackwell, 1981), sec. 225. *Remarks on the Philosophy of Psychology*, vol. 2, trans. C. G. Luckhardt and M. A. E. Aue (Oxford: Blackwell, 1980), sec. 570, contains the closely related formulation without the parenthetical about the doctor. The position is not behaviorism: emotion does not reduce to facial movement, and not every aspect of another's mental life lies open to view. Zahavi's 2011 treatment argues that direct social perception can depend on context and background knowledge without becoming an inference from an intermediary; it also emphasizes that the possibility of mistake and deception does not cancel its experiential character. Vasudevi Reddy, *How Infants Know Minds* (Cambridge, MA: Harvard University Press, 2008), and Shaun Gallagher and Dan Zahavi, *The Phenomenological Mind*, 2nd ed. (London: Routledge, 2012), chs. 8–9, develop the empirical and phenomenological case against treating theory-theory or simulation theory as the universal epistemic basis of social understanding. None of these sources rules out inferential processing or explicit reasoning in difficult cases; the narrower claim is that ordinary warrant need not rest on Millian analogy from one's own case.

**5.** The ELIZA effect takes its name from Joseph Weizenbaum's ELIZA, a 1966 natural-language program whose "DOCTOR" script parodied a Rogerian psychotherapist by reflecting users' statements back as questions. Weizenbaum reported that users, including his secretary, became emotionally involved and wanted privacy with the program despite knowing how simple it was: *Computer Power and Human Reason: From Judgment to Calculation* (San Francisco: W. H. Freeman, 1976), Introduction, 6–7; and "ELIZA — A Computer Program for the Study of Natural Language Communication between Man and Machine," *Communications of the ACM* 9, no. 1 (1966): 36–45. The trigger-and-tracker diagnosis belongs to this book rather than to Weizenbaum. Chapter 25 develops it as a claim about anthropomorphic cue-reading and explains why knowledge of a cue's engineered provenance can screen off the felt presence as independent evidence. The screening is narrower than it sounds, and Chapter 24 states the scope: it rides on an independent inspection of the architecture rather than standing in for one, since a provenance fact by itself disqualifies nothing.

# PART FOUR — Artificial Intelligence

## Chapter 20: Functionalism and Computationalism

> *"We could be made of Swiss cheese and it wouldn't matter."*
>
> — Hilary Putnam, "Philosophy and Our Mental Life," 291

---

CHAPTER OVERVIEW

Functionalism starts with a claim I want to keep: carbon has no prior right to host a mind, and biological tissue is not required. The view goes wrong when it shuts out the world and treats inner cause and effect as the entire mind: matching inner states can reach different things in different worlds, while different states can reach the same thing. No narrow role can fix what a state means, and computational functionalism inherits the mistake when it treats a program state as enough; one program on two kinds of hardware does not prove that both systems share a mental state. As the later Putnam realized, his cheese was only half the story. He left the world out. A wider functionalism can bring in history, place, social use, the body, and links across the whole agent; at that point it begins to look like the world-involving account I defend.

Philosophers call the general theory *functionalism*, and for several decades it came as close to orthodoxy as philosophy of mind allows.[1] Functionalism identifies mental states by the causal roles they play rather than by their material composition. **Computational functionalism**, one influential child of the theory, adds that the relevant roles can be specified as computational states in a program.

Start with a program too simple to confuse anyone. Almost every programmer begins with some version of this one, here in C++:[2]

```
#include <iostream>

int main() {
    std::cout << "Hello, World!" << std::endl;
    return 0;
}
```

Type it, compile it, run it, and the words appear on your screen:

"Hello, World!"

The greeting has an author: *you, the programmer*. Displaying it gives us no reason to posit a new speaker. You might just as well have printed those words on a piece of paper. Nothing in the paper means to say hello to the world. Nobody is home. I offer this bit of common sense as a premise: your reproduction of meaningful words does not by itself establish a new mind.

Now add memory, recursion, and billions or trillions of parameters calculated from machine learning on enough text to fill ten Libraries of Congress.[3] Add photos and videos and other media.

At some point, according to some computational functionalists, a program something like this becomes a mind.

The claim concerns what learning and organization produce: states that retain information, correct one another, guide inference, and alter conduct in new situations. If the right causal organization constitutes a mind in us, why should its realization in a machine count for less? Get that organization right, the argument says, and the mind comes along.

### 20.1 What Functionalism Got Right

For much of the twentieth century, philosophy of mind hung between two unhappy options. The behaviorist treated a mental state as a disposition to behave: believing it will rain amounts to being disposed to carry an umbrella. But no behavior follows from a belief by itself. Up goes the umbrella only if you also want to stay dry, notice the clouds, and have not decided the walk is worth a soaking.

The identity theorist identified each mental state with a brain state: pain with the firing of C-fibers, full stop. Tidier, yes, but at an embarrassing price. A being without C-fibers, whether an octopus, a Martian, or anything whose nervous system took another road, could not feel pain. That discovers nothing about octopuses. It reveals a prejudice about plumbing.

Into this impasse stepped Hilary Putnam, whom you met on Twin Earth in Chapter 16, proving that "meanings just ain't in the head." The man who would drive meaning outside the skull first helped found a theory that seemed to put the mind entirely inside the wiring.[4]

Putnam's founding functionalist move defined a mental state by its *role*, not its material. Functionalism identifies pain with whatever state bodily damage causes, which produces wincing and avoidance and interacts with beliefs and desires in the way pain does, keeping a creature off the sprained ankle and teaching it to mind the stove. Define the state by its place in that causal traffic and you free the mind from any particular substrate. The octopus can feel pain. So could the Martian. So could a machine that realized the relevant role.

David Lewis turned the picture into a method, first in a 1966 paper written at UCLA, then in the Princeton papers that followed. Gather what ordinary psychology says about a mental state, regiment its web of causes and effects, and define the state as whatever occupies that role. Science may then discover different occupants in different beings.[5] Pain in us may have one physical basis, pain in the octopus another, pain in a silicon creature a third.

This possibility became *multiple realizability*: one mental kind might occur in different physical systems.[6] The term ran three claims together. Material anti-chauvinism denies any substance a monopoly on mentality; possible multiple realization says a mental kind could have different realizers; actual recurrence says a mental kind does occur across disparate realizers. Keep the first two and treat the third as a question. Different hardware running one program cannot settle it.

Functionalism leaves the door open to artificial minds by refusing to define mentality in terms of neurons or biological tissue.[7] Nothing that follows closes that door. The argument rejects narrow functionalism, not substrate neutrality. If a system built from silicon, cultured tissue, radio-

linked people, empty beer cans connected by string, or Swiss cheese realized the world-involving organization a mind requires, its materials would not disqualify it.

### 20.2 Why Narrow Functionalism Fails

Material neutrality does not by itself bring the world into the account. A theorist can still draw the entire causal role inside the skin. On this **narrow** reading of functionalism, a mental state gets its identity from the role it plays inside an individual. Signals at the sensory boundary lead to the state; the state interacts with other inner states; behavior follows. Input and output descriptions remain narrow enough for an internal duplicate to satisfy them elsewhere. Call some inner state a "water detector," and the narrow role says only which sensory pattern triggers it and which words come out. The role never says whether water, XYZ, a simulation, or an experimenter with a wire produced the pattern.

Computational functionalism takes the further step of specifying those inner roles as states of a program and identifying those states with mental states. An abstract program specifies formal states and transitions among them. Two implementations in the same formal state must therefore share the same mental state, whatever their surroundings or histories. Narrow functionalism makes internal causal role sufficient; its computational child makes formal program state sufficient.

The child inherits the parent's problem. If internal causal role cannot determine content, formal program state cannot determine it either. A more precise program still leaves the world off the page. The computational version makes the mistake precise enough to test in both directions. Internally identical systems can think about different things because their histories and environments differ. Systems with unlike architectures can think about the same thing because their states reach the same worldly objects by different routes. A taxonomy that joins the first pair and divides the second misidentifies the mental kind. Putnam supplies both demonstrations.

Putnam first made the taxonomy precise: same machine table, same mental type, whatever the hardware. His later work on meaning showed that the taxonomy cuts both ways.

**The same narrow state can carry different content.** Oscar on Earth and his molecule-for-molecule duplicate on Twin Earth may share every internal state. Yet Oscar's "water" reaches $H_2O$, while Twin Oscar's reaches XYZ. Their organization matches; their content differs. No narrow computational type therefore suffices for belief identity.

**The same content can cross different narrow states.** Two neighbors may both believe that many cats live on the street. One learned it by watching the strays. The other heard it from a friend and thinks in a language whose words and inferences do not line up with English. Different states, same content.

A computationalist can still hope to zoom out until some shared type appears. Putnam's real case blocks that hope. Imagine a creature whose brain, or whose machine, updates its beliefs on a different schedule and groups similarities we would never count as the same. It can still be right about the cats. Its table of inner states need not share a single state with ours, or even a space of states that maps onto ours. To collect those unlike organizations as one belief, we already treat

them as reaching the same neighborhood cats. That collecting judgment answers to the world and to the practice of interpretation. It is not a computational type we found first and then named.[8]

In *Representation and Reality*, under the title "Why Functionalism Didn't Work," Putnam drew that moral: a brain considered in isolation cannot supply a state common to all and only the thinkers of a given thought. The machine table had drawn its boundary too tightly.

Here comes the part usually lost when Putnam's recantation gets reduced to a funeral notice. He considered a reply, credited to the philosopher of science Richard Boyd, that widened the machine. **Sociofunctionalism** or **global functionalism** would treat "organisms-cum-environments" as the systems whose relations matter. Inner representations would stand in causal relations to objects, practices, and other speakers. Externalism might force this complication, Putnam wrote, but functionalism was "not yet refuted."[9]

That reply widens functionalism itself; it need not preserve the identification of mental roles with program states. On the **wide** reading, relations to the world partly constitute the mental role. A state may represent tissue damage when it has been recruited to covary with damage and alter consumer activity in ways that protect damaged tissue. A perceptual state presents a tomato as red partly because its system developed and operated among colored surfaces. A speaker thinks about water partly because her word and concept participate in a causal and social history reaching $H_2O$ and her linguistic community. Change those relations while holding the mechanism fixed and the content may change.[10]

Wide roles remain physical and causal. They specify what produces and consumes a state, which conditions its use answers to, how error becomes possible, and how it guides learning and action. History need not return as a present force; it helps constitute what the inner state — the *vehicle* that carries the content — means. The explanation has widened, not departed from nature.

Wide individuation does not make every external object part of the cognitive system. My belief about New York depends for its content on New York; the city does not thereby become one of my cognitive organs. A state can belong to an agent while its identity depends on relations beyond the agent. Tightly coupled tools or environmental loops may sometimes join the realizing system, but that claim requires its own causal argument.

A second caution: a wide role cannot mean "whatever produces mentality." Each addition must earn its place. Tracking history and producer-consumer use help fix content; mode, attention, poising, and integration bear on consciousness; flexible coordination bears on understanding. Calling every relation *functional* does not collapse these questions.

This separates two claims packed into *substrate independence*.

1. **Material neutrality** says composition alone does not settle mentality.
2. **World independence** says an inner profile fixes the mental type regardless of environment and history.

The first opens the door to artificial minds without carrying multiple realization through as a theorem. Putnam's cross-classifications defeat the second.

A separate computational question then remains. **Abstract-program sufficiency** says that the right formal transition structure suffices for mentality. **Computational realization** allows that a concrete, recurrent, historically embedded computational organization may constitute a mind when it realizes the relevant content-bearing, integrative, and phenomenal relations. This book rejects the first and leaves the second open. Computation poses no objection when a system realizes the complete wide organization a mental capacity requires; program identity never establishes that it does. The program alone does not think. A suitably organized computational agent might.

Putnam nevertheless rejected the widened repair. No finite rule, he thought, could capture the open-ended background of beliefs, charitable constraints, community practices, and worldly facts that makes one interpretation better than another. Wide functionalism therefore failed as the reductive identity theory he had wanted.

That failure need not carry naturalism down with it. Putnam asks for more than a procedure that decides hard cases: he asks for a finite definition of intentional relations in physical and computational terms, one that holds across physically possible cases. Rejecting a decision procedure alone would not answer him. The alternative defended here starts with capacity-specific causal and historical explanations. To count as a reduction rather than a redescription, those explanations must specify the relevant relations without presupposing the mental facts they explain. Whether they can cover every possible case remains a further debt, not a victory won by widening the diagram.

Putnam pressed a separate complaint against computationalism. On a permissive account of implementation, he argued, every ordinary open system could be read as running any finite program that receives no outside input. A rock, redescribed cleverly enough, runs the program of your choosing.[11] Chapter 22 takes up that argument and sides with the reply that genuine implementation requires observer-independent causal organization. Grant the non-trivial implementation; Putnam's content point survives. Running a program does not by itself determine what its states concern.

The disagreement with Putnam concerns what his failed universal reduction rules out. I retain the possibility that physically realized organism-world relations constitute content-bearing roles, while accepting that the role framework alone has not explained them. The positive case must come from the accounts of content, understanding, and consciousness, each defended on its own terms.

### 20.3 Why Computationalist Multiple Realizability Fails

Apply those cross-classifications to the computationalist's own argument. The case for multiple realizability ran upward from material to role:

1. Mental states should not be identified with one neural or physical state type.
2. Different physical structures can occupy the same relevant causal role.
3. Therefore a mental type may have different physical realizers.

Machine-state functionalism made the second premise look clean by joining two claims: different physical systems can implement the same computational state, and that state constitutes the mental state. Together they yield the chain:

> Different hardware ⟶ same computational state ⟶ same mental state.

Grant the first arrow. What counts as the same computation depends on how we individuate programs and states, a problem Chapter 22 will inherit. For now, suppose two physical systems non-trivially implement the same program. We have established multiple implementation of a computation.

The second arrow breaks. A shared narrow computational type does not suffice for a shared contentful mental type, and a shared contentful type need not correspond to one narrow computational type. The classifications cross-cut. Multiple physical implementations of one program therefore need not realize one mental state. The hardware-software analogy reaches computation; it does not cross the remaining distance.

The best computationalist can grant all of this and refuse defeat:

> Earth Oscar and Twin Oscar have different broad contents, but the same narrow machinery explains their matching local transitions and proximal behavior. Psychology needs that reusable causal kind. A complete mental state can therefore consist of a narrow computational type plus an external reference-fixing factor. Multiple realizability was always a claim about the narrow causal type, or about that two-factor package, rather than broad content alone.

The concession should stand.[12] A narrow computational description may identify a vehicle, explain local transitions, predict behavior, and characterize domain-specific competence guided by representational content. Computation can remain part of the mechanism after computationalism fails as a theory of mental kinds.

Calling the narrow vehicle a *psychological type* does not make it the state by which a subject thinks about water, perceives damage, or understands a claim. If multiple realizability concerns only that vehicle, it has become multiple implementation of a causal component. If the external reference-fixing factor enters the type, the mental kind has gone wide. Calling the package *wide computation* changes no result: computation remains part of the realization, but sameness of program no longer establishes sameness of mind.

A computational vehicle may run wholly inside the system and cause the next transition. Its content depends on environmental sources, selection or learning history, consumers, and community practices where relevant. One vehicle type may bear different contents in different embeddings; different vehicles may bear the same content when their total systems realize the relevant wide role.

The negative conclusion has an exact scope: multiple implementation of one program does not establish multiple realization of one contentful mental state. A biological system and a silicon system might nevertheless both represent bodily damage. We would need evidence that their mechanisms track damage through the relevant histories, feed consumers organized to protect the body, and guide learning and avoidance. Different materials pose no objection. The shared capacity must be established rather than read off the program.

The boundary of the total base may vary. Sometimes the environment helps individuate a state while the immediate mechanism remains inside the agent. Sometimes a recurrent control loop crosses skin, tool, and environment so densely that the coupled system forms the best candidate realizer. The boundary follows the relations that explain the capacity; wide functionalism cannot settle it by decree.

What remains is material neutrality and an open empirical question. Some world-involving capacities may recur across strikingly different total systems; some finely specified mental kinds may not. Functionalism alone guarantees neither result. "Same program, same mental state" fails. Carbon enjoys no monopoly. Silicon receives no promissory note.

### 20.4 What the Wide View Still Owes

A critic may now suspect that wide functionalism wins by moving the boundary whenever trouble appears. Block's nation, the relabeling objection, and Chalmers's fading qualia are the three places that suspicion still bites. They do not hand the wide view a victory. They show what it still owes.

#### 20.4.1 Block's Homunculus Nation

> Ned Block asks us to build a system whose narrow functional organization duplicates a human brain in pain.[13] Take a large population, give each citizen a radio and a local machine-table task, and have them collectively implement for one hour the transition profile that a human brain instantiates after a stubbed toe.[14] Block connects the organized population by radio to the sensory and motor neurons of an artificial, brainless body, making the population function as its brain. The body closes the cheap escape that the thought experiment forgot perception and action.
>
> Set it running. The inputs match. The outputs match. Every stipulated inner transition matches. Does the *nation* hurt? Is there, beyond the billion busy citizens, a further subject who feels the toe?

The intuition that there is not proves hard to shake. Block calls this an absent-qualia case: the functional wiring appears perfect and the lights appear off.[15] Narrow functionalism must bite the bullet. If the specified internal role exhausts mentality, then the nation hurts, and its strangeness shows only our parochial attachment to familiar materials.

Intuition will not settle the case. If the stipulated facts truly sufficed for consciousness, disbelief would not make them stop sufficing. A wide reply has to name a relation Block did not write into the hour. It does not yet have a verdict about the pain.

Block stipulates an hour of narrow transition matching plus present bodily traffic. At a fine enough grain, that can include current producer-consumer transitions, memory dispositions, learning dispositions, and control of the artificial body. Grant every one of those causal facts. What the text does not stipulate is an address-fixing history for the assembled whole: a past in which those transitions came to reach damage and color, rather than a table copied from a creature whose states already reach them. Present contact is not by itself a representational past.

That gap concerns content, not yet feeling. Whether the nation hurts depends on whether any state of the whole presents that content under a bodily-affective mode, poised and integrated for a persisting subject. Wide functionalism does not answer that question. The identity claim does, and only after the nation's organization has been shown rather than read off its radios.[16] Block's hour leaves the showing undone. It does not prove the lights are off.

The same answer has to live with its own consequence. Suppose the nation-cum-body persisted, learned through its own encounters, developed wide states whose consumers answered to damage and color, and realized the perceptual and bodily-affective organization consciousness requires.[17][18] Citizens and radios would then supply no objection. Wide functionalism should count that system as a candidate mind. A refusal based only on the feeling that a nation is the wrong kind of thing would relapse into the chauvinism Section 20.1 rejected. History matters only where it constitutes the present state's content and role. What the one-hour case still owes is the showing. A prejudice about nations will not do the work.

### 20.4.2 Has Functionalism Merely Changed Its Name?

The second objection presses harder. Either content-fixing relations count as more functional organization, in which case the theory expands until it becomes trivially true, or they count as something beyond function, in which case functionalism has been abandoned. Which is it?

Take the first horn with conditions. The wide view enlarges functional organization. That is no embarrassment if the added relations receive independent descriptions, and if they can fail while the internal table stays fixed. A state tracks damage because of a causal history, not because we first call it a pain. A consumer makes that state answerable to damage because its operations vary with the state in ways selected or learned for dealing with damage. A linguistic state reaches water through demonstrative chains, environmental contact, and social deference. These relations can fail while the internal transition table remains fixed. They predict differences the narrow view cannot see.

The theory would become empty if it defined the relevant role as "the role played by whatever has the mental state." It does not. It offers defeasible, capacity-specific conditions. Content requires a source of correctness conditions. Whole-agent understanding requires flexible coordination among content-guided capacities. Consciousness requires a state to present its content under the right mode and make it poised and integrated. These claims may prove wrong, but they can be tested and criticized separately. A theory with joints can lose an argument.

Nor does width smuggle in dualism. The wide role consists in causal, historical, social, and organizational relations among physical systems. Calling the state relational does not make it immaterial. A key fits a lock partly because of a relation between their shapes; the fit does not float outside nature. A belief fits a world in a more complicated way, but complication does not change its ontological address.

The residue between wide functionalism and this book's world-involving representationalism may indeed come down to filing, on the negative thesis. I am content with that much. Anyone who rejects internal sufficiency, refuses material chauvinism, and explains mentality through the right world-involving organization has crossed the important distance, whatever appears on the folder tab. The cases in this section have not yet drawn the joints. They name relations that would have to fail, while the inner table stayed fixed, before the widened label earns the name of a theory.

### 20.4.3 Organizational Invariance

> David Chalmers presses the hardest remaining case. A property dualist, he nonetheless defends the *principle of organizational invariance*: systems with the same fine-grained functional organization have qualitatively identical experiences.[19][20] His argument asks us to replace a person's neurons one at a time with functionally identical silicon chips. Each chip performs the signaling job of the cell it displaces; the subject continues speaking and acting as before.
>
> Could experience fade while the subject insists that the tomato remains vividly red and the pain remains sharp? That would make the subject radically and systematically wrong about her present experience. Could consciousness persist through thousands of replacements and vanish with one final chip? The boundary looks arbitrary. Chalmers concludes that experience must ride with the organization throughout. If material replacement leaves mentality intact, Block's strange materials cannot by themselves turn the lights off.

Wide functionalism should welcome what the replacement story shows, and stop there. The story preserves far more than an abstract transition table. It preserves the subject's body, environment, practical engagements, social setting, and causal history. Each chip enters an already functioning, content-bearing economy and inherits the role of the cell it replaces. Mentality can stay through that change of materials. Replacement shows that the materials can change. It does not show that we can ignore the world.[21]

Chalmers's principle is not limited to the person whose neurons are being replaced. Any system with the same fine-grained functional organization is to have qualitatively identical experiences. Fading and dancing qualia draw their force from a threat of systematic error: if experience faded or danced while organization and the subject's judgments stayed fixed, she would be wrong about her present experience and would not notice. The replacement story never creates that threat.

It holds the wide organization fixed along with the judgments. A fresh duplicate, matching the inner topology and the reports while lacking the subject's history, is a different case. To say that matching transitions do not establish the same content is to deny the principle. It is not yet an answer to the argument.

Separate three cases.

1. A replacement descendant inherits the original subject's content-fixing history and complete phenomenal organization. Material substitution can leave mentality in place. That case is settled here.
2. A freshly assembled narrow duplicate matches the internal topology, judgments, and utterances. Whether fading-qualia reasoning still binds that duplicate, once history is withheld, is what the wide view still owes.
3. A differently built system may independently acquire an equivalent history and wide organization. That possibility needs its own evidence. A shared program does not supply it.

If Block's nation acquired an equivalent history and integration, the difference would narrow, and the wide functionalist should grant the capacity. The theory cannot use history as a ceremonial password, denying mentality forever to an artifact because it did not grow in the approved way. History matters only where it constitutes the present state's content and role. It cannot, in this section, stand in for the answer Chalmers's unnoticed-change argument still requires.

### 20.5 The Background No Rule Contains

Putnam reaches narrow computationalism from the side of meaning. Hubert Dreyfus reaches it from the side of action.

In the early 1960s, Dreyfus taught philosophy at MIT while its young artificial-intelligence project pursued the prospect that engineers might explain the mind by programming one. His objection concerned something ordinary expertise makes easy to overlook: knowing what matters in a situation.

The wall is *relevance*. To act intelligently in a real situation, a system must treat some available facts as bearing on what to do and ignore the boundless rest. A program that applies explicit rules to represented facts faces an awkward question: which facts? Most truths about any situation do not matter to the task. Give the program a rule for selecting the relevant ones. That rule helps only if the system can tell when *it* applies. If that judgment must itself come from another explicitly represented rule, the regress begins one floor up.

A cook in an unfamiliar kitchen does not run the regress. The pot, the flame, and the moment to lower the heat show up as relevant against a background of embodied involvement she acquired without writing it into propositions. Her know-how does real work. Trying to spell all of it out creates more facts whose relevance would itself need settling. AI met a close relative of the difficulty as the *frame problem*. Dreyfus treated it not as one missing rule but as evidence that the rule-and-representation picture had omitted the background against which any rule matters.[22]

The criticism lands against classical, narrow computationalism. A finite stock of inner representations and explicit rules does not manufacture relevance by adding another representation and

rule. But Dreyfus does not thereby refute a wide, world-involving account. Such an account may treat skilled coping as learned, embodied, holistic, and dynamical. It can explain relevance through a history in which some distinctions became action-guiding within recurrent organism-world traffic. No homunculus consults a relevance list. The organization has been shaped so that some possibilities solicit response and others recede.

Modern learning systems make that reply concrete. Training can give mechanisms causally potent, task-shaped states with correctness conditions; recurrent perception and action can continually revise them. Natural selection and gradient-based learning differ profoundly, but both remind us that a function may depend on history rather than an explicit rule stored inside the present token.[23] The remaining questions concern what those states represent, how flexibly their consumers use them, and whether the mechanisms compose one persisting agent. Learning answers none automatically. It changes the right question from "Where is the relevance rule?" to "What organization did the history produce?"

I therefore take Dreyfus's wall and leave his blueprint. Dreyfus often drew an anti-representational moral: intelligent coping does not proceed by manipulating inner content. This book keeps representation and asks how world-involving histories give its states correctness conditions.[24] A wide naturalistic account can keep both the content and the background by locating them in an embodied, temporally extended organization rather than a snapshot of rules.

Putnam and Dreyfus converge on a narrower verdict than either sometimes drew. Formal transitions alone do not supply meaning or relevance. A running computational mechanism can still participate in historically grounded, content-sensitive loops; several such mechanisms can constrain one another in flexible action. The cook knew the diagram was incomplete. She did not prove that no diagram could ever include the kitchen.

### 20.6 The Door to the Room

Chapter 21 tests a different question: whether formal processing establishes understanding by the implemented whole. Its thought experiment can block that shortcut without disqualifying a historically embedded system that acquires content and integrates it across a cognitive life.[25]

### Chapter Summary

**The claim.** Carbon and neurons hold no monopoly on mind. Material make-up alone cannot settle mentality. Yet a narrow functional role cannot fix content, because a causal diagram cut off from the world cannot determine what its states concern.

**Where the argument stands.** Putnam's externalism blocks the inference from one narrow program state to one contentful thought. Computational kinds still explain mechanisms; a historically embedded computational system might also realize a mind. Multiple implementation of one program does not establish multiple realization of one mental state. Material replacement can preserve a state when the subject's wider organization and history remain fixed. Block's hour, and a fresh duplicate without that history, remain open.

**What's left open.** Content-bearing organization does not by itself settle understanding, state consciousness, or subjecthood. It also does not settle autonomy or welfare. Which implemented systems meet those separate conditions remains an empirical question.

**The hand-off.** Remove the world and functionalism becomes syntax with excellent wiring. The clerk in the Chinese Room waits to show how little that guarantees.

**Footnotes**

**1.** Block opens "Troubles with Functionalism" by observing that functionalism "may now be dominant," then immediately notes that the label covers several distinct projects — reformulations of behaviorism, mind-machine analogies, applications of empirical psychology, and arguments about mind-brain identity. The version pressed and answered here is the strong one: each mental state is *identical* to a functional state, a state defined by its causal relations to inputs, outputs, and other inner states. Computational functionalism adds that the relevant functional states are computational states and that running the right program suffices for mentality. Nothing in the broader argument against narrow functionalism turns on machine-table formalism specifically; the later computational argument does. Ned Block, "Troubles with Functionalism," in *Perception and Cognition: Issues in the Foundations of Psychology*, ed. C. Wade Savage, Minnesota Studies in the Philosophy of Science, vol. 9 (Minneapolis: University of Minnesota Press, 1978), 261–325, esp. 261–63.

**2.** The example goes back to Brian Kernighan's internal Bell Laboratories memorandum *A Tutorial Introduction to the Language B* (1972), where it illustrates external variables; the phrase arrives split across three of them because a character constant in B held at most four ASCII characters. Kernighan carried it into C in a second Bell Labs memorandum, *Programming in C: A Tutorial* (1974), and it entered the discipline's common stock through Brian W. Kernighan and Dennis M. Ritchie, *The C Programming Language* (Englewood Cliffs, NJ: Prentice-Hall, 1978), sec. 1.1, which prints it in lower case and without terminal punctuation — "hello, world". The first edition's program carries no include directive; that line was added in the second (1988). The C++ form given above follows later convention, and nothing in the argument turns on the language, capitalization, or particular string. Chapters 22 and 23 distinguish implementation from realization and formal structure from world-grounded content.

**3.** The premise commands wide assent, though not always for the same reasons. Panpsychism and the phenomenal versions of Russellian monism place phenomenal properties at the fundamental level rather than having them arrive with a particular arrangement of matter; whether combining them can ever make a new subject is their combination problem, and none of them claims that running a short greeting program does so. Russellian views of the protophenomenal kind are a separate case. Chapter 10 takes up that family in detail. Integrated information theory ties experience to intrinsic cause-effect power and, on its proponents' own assessment of conventional computer architectures, finds very little of it there. I assume the premise here in its plain form: writing and running this program brings no new subject into being. Appendix B.18 engages integrated information theory in full. Recent disclosed pretraining counts supply the scale behind the comparison: DeepSeek-V3 used 14.8 trillion tokens, Llama 3.1 405B used 15.6 trillion, and Qwen3 used 36 trillion. DeepSeek-AI, "DeepSeek-V3 Technical Report," arXiv:2412.19437 (2024);

Abhimanyu Dubey et al., "The Llama 3 Herd of Models," arXiv:2407.21783 (2024); Qwen Team, "Qwen3 Technical Report," arXiv:2505.09388 (2025). A widely repeated 1997 estimate put the text of the Library of Congress's books at about 10 terabytes. Using the rough English conversion of four characters per token yields approximately 2.5 trillion tokens per Library and places those disclosed corpora at roughly six to fourteen Libraries; the main text uses ten as a midpoint illustration. See Leslie Johnston, "How Many Libraries of Congress Does It Take?," *The Signal*, March 23, 2012; OpenAI, "Understanding and Counting Tokens." The comparison measures token-scale exposure, not unique books: training data may include repeated material, code, synthetic text, and many languages, and many leading closed-model developers disclose no comparable total.

**4.** Hilary Putnam, "Philosophy and Our Mental Life," in *Mind, Language and Reality: Philosophical Papers, Volume 2* (Cambridge: Cambridge University Press, 1975), 291–303, at 291 (epigraph); and "Psychological Predicates," in *Art, Mind, and Religion*, ed. W. H. Capitan and D. D. Merrill (Pittsburgh: University of Pittsburgh Press, 1967), reprinted as "The Nature of Mental States" in *Mind, Language and Reality*, 429–440, esp. 432–37. Putnam models the whole organism as a probabilistic automaton whose functional state is fixed by sensory inputs, motor outputs, and transitions among states rather than by physical composition. Functionalism itself remains compatible with dualism. The lesson retained here is that material labels alone do not decide mentality, not that cross-material recurrence follows automatically. Putnam later abandoned the reductive project: see *Representation and Reality* (Cambridge, MA: MIT Press, 1988), esp. ch. 5.

**5.** David Lewis, "An Argument for the Identity Theory," *The Journal of Philosophy* 63, no. 1 (1966): 17–25; "Psychophysical and Theoretical Identifications," *Australasian Journal of Philosophy* 50, no. 3 (1972): 249–258, esp. 249–56; and "Mad Pain and Martian Pain," in *Readings in Philosophy of Psychology*, vol. 1, ed. Ned Block (Cambridge, MA: Harvard University Press, 1980), 216–222. Lewis's Ramsey-sentence method defines mental terms as whatever occupies the causal roles specified by the platitudes of common-sense psychology, allowing the occupant to differ across beings (the "mad pain" and "Martian pain" cases dramatize the two ways role and realizer can come apart). The device makes precise the sense in which functionalism keeps the role fixed while leaving the realizer open.

**6.** Putnam's original physical multiple-realization argument appears in "The Nature of Mental States," 435–37. In *Representation and Reality*, introduction, xii–xv, and ch. 5, esp. 74 and 82–84, he later argues that the same belief can occur across unlike physical states and across unlike, even nonmappable, computational state spaces. The later result defeats multiple realizability as computationalists had used it: physical multiple implementation of one narrow computational type does not establish multiple realization of one mental type. Whether one world-involving type recurs across different total physical bases remains a further question.

**7.** Functionalism entails neither physicalism nor computationalism. A dualist could identify mental kinds by causal role while holding that some nonphysical property occupies that role; a functionalist need not identify the relevant roles with states of a program. Its importance for strong AI lies elsewhere: nonbiological systems remain possible occupants of mental roles. The chapter therefore keeps material neutrality while arguing that an internally specified organization

leaves out the world-relations that fix content. It does not infer a shared mental type merely from two realizations of one program.

**8.** The argument rests on meaning holism together with the division of linguistic labor and the environmental determination of reference: Putnam, "The Meaning of 'Meaning,'" in *Mind, Language and Reality*, 215–271, taken up in Chapter 16 of this book. *Representation and Reality*, ch. 5, esp. 74 and 82–87, draws both halves of the lesson. Subjects can share a belief while occupying different, even nonmappable, computational state spaces, so no particular narrow computational type is necessary for that content. Twin-Earth counterparts can share their narrow organization while differing in reference, so no such type is sufficient. Putnam considers collecting unlike states into an equivalence class but argues that the relation doing the collecting is interpretive rather than computable: correct interpretation answers to the world and speech community, not merely to a formalism internal to the believer. His "cats in the neighborhood" example illustrates the point.

**9.** Hilary Putnam, *Representation and Reality* (Cambridge, MA: MIT Press, 1988), ch. 5, "Why Functionalism Didn't Work," 73–90, esp. 74–75 under "Sociofunctionalism." Putnam credits Richard Boyd with the reply, lets the relevant system include society and environment—"organisms-cum-environments"—and concludes at that stage that externalism has made functionalism more complicated but "not yet refuted." In ch. 6, esp. 92–94, he denies that a universal computable equivalence relation can collect diverse computational states into belief types; his closing diagnosis adds an internal-realist objection to the ontology presupposed by wide reduction. The retrospective essay in *Words and Life*, ed. James Conant (Cambridge, MA: Harvard University Press, 1994), 441–59, esp. 445–46, calls the widened proposal "Global Functionalism." Putnam did not endorse it. His reduction requirement in *Representation and Reality*, 77–80, concerns finite definability in physical/computational vocabulary across physically possible cases, not merely an effective decision procedure. The chapter accepts the externalist constraint and pursues capacity-specific physical reductions without claiming that this strategy already answers his universal challenge or accepting his later internal-realist conclusion.

**10.** The name and the canonical statement come from Gilbert Harman, "Wide Functionalism," ch. 13 of *Reasoning, Meaning, and Mind* (New York: Oxford University Press, 1999), 235–43. Harman treats psychological explanation as a kind of functional explanation whose relevant functions concern perceiving and acting in relation to the environment, and takes the wider functionalism to be more plausible than its narrow, more solipsistic rival. He does not use the phrase "long-arm." That label is Ned Block's, in "Inverted Earth," *Philosophical Perspectives* 4 (1990): 53–79, for the distal roles Block takes from Harman — roles whose input side reaches past the sensory surfaces to conditions in the world. The book takes up Inverted Earth in Part Two alongside the identity claim. Putnam's earlier widening (note 9 above) extends the system to organisms-cum-environments; Harman widens the role itself. Both moves meet the same externalist constraint from different directions: the individuating relations cross the skin.

**11.** *Representation and Reality*, appendix, esp. 120–25, argues that, on an unconstrained notion of implementation, every ordinary open system realizes every inputless finite-state automaton, so computational description by itself ascribes almost nothing and cannot constitute mentality.

David J. Chalmers, "Does a Rock Implement Every Finite-State Automaton?," *Synthese* 108, no. 3 (1996): 309–333, secs. 6 and 9, replies that genuine implementation requires a system's causal organization to mirror the automaton's combinatorial state-transition structure, an observer-independent constraint the triviality construction fails to satisfy. Section 6 introduces that constraint; section 9 restates it. Section 7, "Open questions," still calls the account "useful but not perfect" and leaves certain false implementations open. This book follows Chalmers on that question; Chapter 22 treats observer-relativity directly. The concession leaves *narrow* functionalism no better off: a system may run a program nontrivially while its internal organization still fails to determine what its states are about.

**12.** The retreat packages a narrow computational type with an external reference-fixing factor. Jerry A. Fodor, *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (Cambridge, MA: MIT Press, 1987), ch. 2, treats narrow content as a function from contexts onto truth conditions (47–53). Ned Block, "Advertisement for a Semantics for Psychology," *Midwest Studies in Philosophy* 10 (1986), identifies the internal factor with conceptual role. Putnam, *Representation and Reality*, ch. 3, "Fodor and Block on 'Narrow Content,'" 43–56, argues that meaning holism and the division of linguistic labor "militate against the whole division of meaning into a 'narrow content' and a 'broad content'" (55). The present reply grants the reusable causal kind and denies that it is the mental type whose multiple realization was at issue.

**13.** Block has pressed the same conviction — that reductive theories of mind keep omitting something real — across absent qualia, the distinction between phenomenal and access consciousness, and the "mental paint" objection to representationalism. The homunculus-nation is the version aimed squarely at functionalism's sufficiency thesis. Block, "Troubles with Functionalism" (1978), esp. 279–81.

**14.** Block's "homunculi-head" and "Chinese nation" (or "China-body system") cases, both from "Troubles with Functionalism" (1978, esp. 279–81). The scenarios are built so that the system's functional organization is, by stipulation, identical to a human's, isolating the single question of whether that organization *suffices* for phenomenal consciousness. In the China case, roughly one billion participants carry out local machine-table tasks by radio while the nation controls an artificial body. Block chose the approximate population by reference to the estimated number of neurons in a human brain; he did not assign one citizen to each neuron. The setup makes vivid that every narrowly specified functional fact can be satisfied while the materials and their arrangement are nothing like a brain's.

**15.** Block distinguishes *absent qualia* (the system satisfies the functional specification but has no phenomenal experience at all) from *inverted qualia* (two systems functionally identical but with systematically swapped experience). A parallel pressure comes from inversion, but I leave it aside here: turning the inverted spectrum against a *representational* theory of consciousness, rather than against functionalism, opens a separate and harder debate, taken up in Part Two's treatment of the identity claim. Against functionalism's sufficiency thesis specifically, the absent-qualia case is enough. For the standard taxonomy see Michael Tye, "Qualia," *The Stanford Encyclopedia of Philosophy* (Fall 2025 Edition), ed. Edward N. Zalta and Uri Nodelman, sec. 4, substantive revision September 19, 2025.

**16.** The inference is deliberately epistemic. Block stipulates a narrow organization but not the complete world-involving intentional structure. His case therefore does not establish content or consciousness; neither does this reply establish their absence. On the identity claim defended in Part Two, absence of the requisite world-presenting structure would entail absence of the corresponding phenomenal character, but whether the nation has that structure must be shown independently rather than read off its unusual materials.

**17.** This is semantic externalism brought to bear on the theory of consciousness: what a state represents is not settled by anything internal to the system at a time but by its causal-historical relations to the environment, so two internally identical systems can differ in what their states are about because they are embedded in different worlds. This is precisely the resource narrow functionalism lacks: it specifies internal role exhaustively and is silent about the world-relations that fix content. The positive account is developed in Part Three; see especially the chapters on teleosemantics and on semantic externalism.

**18.** The identity claim defended in Part Two begins from Tye and Dretske but belongs to the book's cumulative argument. It identifies phenomenal character with an experience's complete world-presenting intentional structure: independently fixed content under a perceptual or bodily-affective mode, with determinacy, field structure, and attentional organization. Michael Tye, *Ten Problems of Consciousness* (Cambridge, MA: MIT Press, 1995), esp. ch. 5, sec. 5.2; Fred Dretske, *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995), esp. ch. 3. Tye also distinguishes narrow functional duplication from "maximal" or wide duplication, which preserves external causal interactions, present and historical, and argues that maximal functional duplicates cannot differ phenomenally; *Ten Problems*, ch. 7, 196–201. The book treats history as content-fixing and uses a separate distributed profile of recurrence, availability, flexible influence, and control to assess state consciousness. Neither history nor that profile gets added to phenomenal character after the fact. The inference here runs from absence of the requisite world-presenting structure to absence of the corresponding phenomenal character; it does not run from content alone to consciousness.

**19.** David J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press, 1996), esp. ch. 7. The argument has unusual dialectical force because Chalmers rejects the logical supervenience of consciousness on the physical yet independently defends organizational invariance. The case for invariance therefore does not merely assume reductive functionalism.

**20.** Chalmers, *The Conscious Mind*, ch. 7, "Absent Qualia, Fading Qualia, Dancing Qualia." The principle of organizational invariance holds that "given any system that has conscious experiences, then any system that has the same fine-grained functional organization will have qualitatively identical experiences" (249). The fading- and dancing-qualia arguments support it via gradual neural-replacement scenarios: a subject whose qualia faded or danced while functional organization held constant would be radically and systematically mistaken about her own present experience despite being fully rational. Chalmers judges that hypothesis "significantly less plausible" than the hypothesis that rational conscious beings are generally right about their experiences; he offers it as a strong case against such dissociation in nature, a claim about natural rather than logical possibility, not as a proof that the alternative is incoherent.

**21.** The chapter does not claim to have refuted Chalmers's universal principle. His neural-replacement scenarios preserve the subject's body, environmental traffic, causal history, and perceptual organization while changing material. They support continuity in that lineage, and only that: a replacement descendant inherits broad content because its history remains fixed. They do not show why fading- or dancing-qualia reasoning would fail for a freshly assembled duplicate that matches internal topology, judgments, and reports while lacking that history. A differently built system may qualify if it independently acquires an equivalent wide role, but that possibility requires its own evidence. Block's artificial body prevents the cheap claim that his nation lacks present traffic; the narrower point is that the one-hour case stipulates no address-fixing history for the nation as an assembled whole. Any verdict about whether the nation feels waits on the identity claim in Part Two, not on wide functionalism alone.

**22.** Hubert L. Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason* (Cambridge, MA: MIT Press, 1992), the revised edition of *What Computers Can't Do* (New York: Harper & Row, 1972), esp. parts II–III, chs. 5–9; for the relevance regress specifically, part III, ch. 8, esp. 256–59. The relevance regress — that picking out the relevant facts requires a rule, whose application requires a further rule, without end — is Dreyfus's central charge against the rule-and-representation paradigm of "good old-fashioned AI." A close relative appears in what AI came to call the *frame problem*, posed in John McCarthy and Patrick J. Hayes, "Some Philosophical Problems from the Standpoint of Artificial Intelligence," in *Machine Intelligence 4*, ed. B. Meltzer and D. Michie (Edinburgh: Edinburgh University Press, 1969), 463–502, and given its classic philosophical statement in Daniel C. Dennett, "Cognitive Wheels: The Frame Problem of AI," in *Minds, Machines, and Evolution*, ed. Christopher Hookway (Cambridge: Cambridge University Press, 1984), 129–150. Dreyfus's positive account of the holistic background of skilled coping draws on Maurice Merleau-Ponty, *Phenomenology of Perception* (1945), esp. part I, ch. 3, and Martin Heidegger's analysis of the ready-to-hand in *Being and Time* (1927), esp. div. I, ch. 3, secs. 15–18; the "know-how vs. know-that" contrast is Gilbert Ryle's.

**23.** Hubert L. Dreyfus, "Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian," *Philosophical Psychology* 20, no. 2 (2007): 247–268, esp. secs. 1–5. Dreyfus argues that dropping hand-coded rules does not by itself supply absorbed, affectively significant involvement. The present chapter narrows the inference. Training may yield causally internalized content in a mechanism. What does not follow automatically is a unified agent that learns, infers, resolves conflicts, and acts through those contents. Chapter 24 develops that further attribution condition and treats self-maintenance separately as a candidate basis of autonomy and stake.

**24.** The divergence is deliberate. Dreyfus reads the background and relevance regress as reasons to abandon internal representation, a reading developed with reservations in Michael Wheeler, *Reconstructing the Cognitive World* (Cambridge, MA: MIT Press, 2005), 187–189, 222–223. This book keeps representational content and grounds it through selection, learning, consumer use, and world-involving history. For the contrast between Searle's representational account and Dreyfus's content-free background, see Teodor Negru, "Intentionality and Background: Searle and Dreyfus against Classical AI Theory," *Filosofia Unisinos* 14 (2013): 18–34, esp. 25–30. Stake may explain why possibilities matter to an autonomous agent; it supplies no semantic bridge.

**25.** John Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3, no. 3 (1980): 417–424, esp. 417–19, and "Is the Brain a Digital Computer?," *Proceedings and Addresses of the American Philosophical Association* 64, no. 3 (1990): 21–37, esp. 25–29. The chapter adopts Searle's insufficiency objection—formal syntax alone does not yield semantics—while rejecting his biological-naturalist conclusion. The proposed missing relation is substrate-neutral causal-historical grounding in an environment.

## Chapter 21: The Infamous Chinese Room

> *"A physical symbol system has the necessary and sufficient means for general intelligent action."*
>
> — Allen Newell & Herbert A. Simon, "Computer Science as Empirical Inquiry: Symbols and Search"

---

CHAPTER OVERVIEW

Put a clerk inside the Chinese Room and give him rules for moving signs he cannot read. His ignorance proves one thing: he does not know Chinese. I argue that the room's smooth replies prove no more about the program or the larger system, since rules and fluent output leave the real issue open. The systems reply rightly shifts our gaze from the clerk to the implemented whole, which may do what none of its parts can do alone. There, content would need to guide memory, error checks, sight, thought, plans, and acts within one lasting system. A machine with that wider history and organization might understand, so the Room blocks two shortcuts without barring every mind made by a computer.

> Let me describe a system. It receives inputs in Chinese, applies elaborate rules to Chinese symbols, and produces contextually apt Chinese replies. To a native speaker, the answers look indistinguishable from those of someone who understands: they remain correct, responsive, and sensitive to the conversation. From the outside, the system passes.
>
> Now look inside. A person, the operator, sits among buckets of Chinese symbols with a rulebook written in English. When a string arrives through the slot, he consults the book, selects the prescribed symbols, combines them in the prescribed order, and sends them back. He follows every rule meticulously. His answers satisfy the stipulated test. And he understands not one word of Chinese.
>
> That is John Searle's Chinese Room, and in its original 1980 form it generated one of the most sustained debates in the philosophy of mind.

Searle aimed the Room at a live achievement. Roger Schank's group at Yale had built programs that used *scripts*, structured accounts of familiar situations, to answer questions a story left unstated. In Searle's example, a man orders a hamburger, receives it burned to a crisp, and leaves angrily without paying. Did he eat it? The program answers no. It supplies an inference, not a sentence retrieved from the story. Searle disputed the further claim he called *strong AI*: that "the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states."[1] He expressly declined to attribute that thesis to Schank himself. He had no objection to *weak AI*, the use of

computers as tools for studying minds. His objection concerned the stronger claim that the right program already constitutes understanding.

My conclusion is narrower than Searle's. Syntax alone does not suffice for semantics, but the operator's ignorance cannot settle what the whole understands. The Room exposes the need to distinguish those questions; the book's case against formal sufficiency also depends on the world-involving account developed in Part Three. Artificial understanding remains possible.

### 21.1 The Argument in Its Strongest Form

The bounded lesson I draw concerns two different bearers. Operator-level rule-following does not establish the operator's understanding, while output parity supplies evidence without constituting understanding by the implemented whole. Selection, learning, and world-involving history may fix what the whole's states answer to. Their coordinated use may support understanding. The Room supplies no independent argument against an implementation that realizes both.

Chinese symbols already possess public content through the practices of Chinese speakers. The stipulation sets them moving by rule; it leaves open whether their contents guide the whole as an integrated system. That gap in the description is not evidence that every possible room-sized implementation must lack understanding.

### 21.2 The Standard Replies, Each in Its Strongest Form

The standard replies have followed the argument from the day it appeared, and each deserves a full hearing.

#### 21.2.1 The Systems Reply

> Most readers reach for this reply within a minute of meeting the Room. Of course the *operator* doesn't understand Chinese; nobody said he did. But the *system as a whole* (operator, rulebook, buckets, and room) might. On this reply, the argument points at the wrong level of organization.

Searle answered with a single adjustment to the thought experiment: imagine the operator memorizes the entire rulebook and carries it all in his head. Now there is no room, no rulebook, no buckets — just a person with the rules internalized. He walks out into the world and answers questions in Chinese as fluently as before. Does he now understand Chinese?

Memorizing the rulebook saves space. It does not settle who understands. A systems advocate may still say that a Chinese-processing organization runs beneath the operator's awareness. The answer must therefore ask what the implemented whole can integrate and do.[2]

Readers convinced that the argument was answered in 1980 usually hold a stronger modern form. Call it the virtual-mind version: the operator stands to the understanding as hardware stands to software. When the Room runs, a *virtual* Chinese speaker comes into existence, implemented in pencil-work as a chess program gets implemented in silicon. Asking whether the operator

understands Chinese then resembles asking whether the silicon enjoys chess. Substrate ignorance misses the candidate.[3]

Here the virtual-mind reply correctly relocates the candidate. A whole can do what none of its parts does alone. Naming a virtual speaker does not establish its capacities, but neither does the operator's failure to notice that speaker disprove them.

A modern cousin hits harder: no neuron understands Chinese, yet a brain built from neurons does. "No part understands" therefore shows nothing. Any disanalogy between brain and Room must concern organization.

The brain-body system offers evidence of content-bearing states coordinating a continuing life. To compare the Room fairly, we need to know whether its implemented organization does the same. If it does, the operator's ignorance supplies no remaining veto.

### 21.2.2 The Robot Reply

Next, put the Chinese Room inside a robot. Give it cameras for eyes, microphones for ears, motors for limbs, and the capacity to learn through action. Let experience correct its errors and reshape its internal organization; let perception feed memory, planning, and control. Now the symbol manipulation participates in a closed loop with the world. *Now* the system understands.

Eyes and hands change where the signals come from and what the system can do in response. Over a suitable learning history, those loops may fix genuine content. A camera added to an otherwise unchanged rule-following device would not settle understanding; a robot that uses what it learns to correct later perception and action presents a stronger case.[4] The change concerns the realized organization, not simply more input.

### 21.2.3 The Brain Simulator Reply

The brain-simulator reply raises the stakes. Imagine the room contains not a rulebook but a complete simulation of a Chinese speaker's brain — every neuron, every synapse, every firing pattern. The operator follows rules that mirror the causal structure of an actual Chinese-speaking brain. *Surely* that simulation understands Chinese. The intuition deserves respect: if a perfect duplicate of an understanding brain's activity doesn't understand, what would?

No answer follows from the word *simulation*. Reproducing only a formal pattern of neural activity does not settle which of the brain's relevant causal powers the implementation realizes. A model of digestion does not digest merely because its variables mirror a stomach; a physical emulation that transformed matter and extracted energy would. The analogy poses the realization question; it does not decide which causal organization understanding requires. The implementation has to be inspected, not dismissed by label. If it realizes the world-sensitive, temporally extended integration that supports understanding, calling it a simulation does not take that achievement away.

A companion argument of Searle's from 1990 sharpens the challenge: computation itself, he argued, is observer-relative; syntax is not intrinsic to physics, and whether matter counts as running a program is something an observer reads onto it.[5] Chapter 22 grants the strongest answer to him: disciplined, intervention-supporting causal-computational organization can belong

intrinsically to a physical system. That structural victory still leaves the present question open. Does this implementation merely realize an intrinsic computation, or does it also realize content-bearing mechanisms integrated into a cognitive system? A perfect formal model cannot answer by stipulation, and a genuine realization cannot be excluded by vocabulary.

### 21.2.4 The Combination Reply

The original exchange saved its strongest card for fourth. The combination reply, from Berkeley and Stanford, takes the first three replies together. As Searle stated it for its advocates: "Imagine a robot with a brain-shaped computer lodged in its cranial cavity, imagine the computer programmed with all the synapses of a human brain, imagine the whole behavior of the robot is indistinguishable from human behavior, and now think of the whole thing as a unified system and not just as a computer with inputs and outputs. Surely in such a case we would have to ascribe intentionality to the system." Alone, each reply might fall short, the advocates granted; together they become "collectively much more convincing and even decisive."[6]

Give the package its due: it leaves the Room nothing cheap to say. The whole unified system now stands as the candidate. Symbols arrive through sensors and leave through motors; the computation mirrors the synapses themselves. The robot differs from the Room in every way that seemed to matter. A reader who feels the case's pull has noticed the additions, not forgotten the original argument.

What carries the weight? Disconnecting the robot's motors and supplying inputs from a virtual environment would not cleanly subtract its world. The environment would itself have a physical implementation, and the robot could retain a content-fixing history acquired before the change. If understanding survived, it might survive through those relations. That would establish understanding in a different realized system, not sufficiency of an abstract program apart from realization. If understanding failed, we would still need to identify which relevant relation the change removed. The subtraction cannot decide the case in advance.

Searle answered in a different register. We would find it "rational and indeed irresistible" to ascribe intentionality to such a robot, he granted, "as long as we knew nothing more about it"; but "as soon as we knew that the behavior was the result of a formal program, and that the actual causal properties of the physical substance were irrelevant we would abandon the assumption of intentionality." Learning the architecture should change our evidence. What observers would assume and later withdraw, however, cannot by itself settle what constitutes the robot's capacities. The live question concerns what the robot has.

On the account defended here, the combination reply could succeed. If the robot's history fixes content and its machinery integrates that content in flexible, world-sensitive use, nothing in the Room bars understanding. Whether all those relations admit a computational description remains a further question. Their realization, rather than the operator's introspection or our first impression of the robot, must decide the attribution.

### 21.2.5 The Other Minds Reply

> The last reply turns the tables on the skeptic. How do you know any *other person* understands Chinese? You judge from behavior. The Chinese Room produces the right behavior. By your own standards, then, you should attribute understanding to the room.

The reply rightly demands consistent standards of evidence. Fluent behavior can support an attribution in either case. But how we *know* someone understands differs from what understanding *consists in*. Evidence can favor a theory of that constitution without becoming the capacity it supports. We must neither dismiss the Room's performance nor treat a stipulated match in output as a complete theory of understanding.

Chapter 19 denied that ordinary recognition of another mind begins as an inference from neutral behavior. We see grief in a face, attention in a glance, and understanding in a reply made within a world we share. That direct perception remains defeasible rather than magical. Shared embodiment, history, and situation normally bind the expressive cue to a persisting subject; known architecture can also defeat the attribution. The Chinese Room isolates the cue while withholding the wider organization. Its behavior may justify an initial attribution without constituting what would make the attribution true.

## 21.3 From the Room to the Model

A language model replaces the operator's rulebook with learned weights and scales the formal work beyond recognition. Training yields fluency, abstraction, code, and transfer beyond practical hand execution. Mechanisms can also earn original content about word-use distributions where training and use fix accuracy independently of an interpreter. These achievements leave the Room's question open. A paralyzed coach can understand swimming without swimming, and a speaker can refer to Napoleon without meeting him; personal acquaintance is no universal test. The stronger issue is whether inherited or acquired contents enter one persistent cognitive economy that integrates perception, memory, correction, reasoning, and action. In a bare text-only model, considered apart from persistent memory, tools, and world-coupled scaffolding, that integration has not been shown. Mechanism content is real; whole-model understanding does not follow.[7]

Nor does the word *emergence* settle the difference. Scale can produce genuinely new capacities, sometimes abruptly.[8] The question is what organization the scaling produces. Several content-guided capacities that correct and constrain one another may support partial understanding. Broader, persistent, world-involving integration requires further evidence; it does not follow from scale alone. Chapter 23 takes up indirect and multimodal grounding.

## 21.4 What the Room Does Not Show, and What Would Pass It

The Room's valid work ends with two blocked shortcuts: neither an operator's rule-following nor fluent output establishes understanding by the implemented whole. Whether a machine

understands now depends on its content-bearing history and integrated realization, which the thought experiment does not inspect.

**Chapter Summary**

**The claim.** The Chinese Room blocks two shortcuts. The operator's rule-following does not establish Chinese understanding, and fluent output does not establish that the whole system understands. Yet the Room does not prove that no computational whole could understand.

**Where the argument stands.** The Room supplies public Chinese content and formal processing, but it does not show one flexible system integrating that content. The systems and virtual-mind replies thus move the question to the implemented whole, where the operator's ignorance settles nothing. Robot and combination replies add perception, action, learning, and memory; brain simulation still needs realized powers, and current-model fluency still leaves the subject and its organization in question.

**What's left open.** A richer system may link content across one lasting cognitive life that answers to the world. State consciousness, autonomy, interests, and welfare would each need separate evidence.

**The hand-off.** The Room leaves the whole's understanding open. The case against syntax alone also relies on Part Three. The next chapter asks what a running implementation really realizes.

**Footnotes**

**1.** John Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3 (1980): 417–424. The paper was published with a substantial collection of peer commentaries and Searle's replies, making it one of the most thoroughly debated papers in the philosophy of mind. Six named replies are addressed in the original exchange — systems, robot, brain-simulator, combination, other-minds, and many-mansions, the last urging that some future technology might supply the right causal powers; Searle agrees it might and observes that the concession abandons strong AI's defining claim, that programming alone already suffices. For a useful overview of the subsequent debate, see Ned Block, "The Mind as the Software of the Brain," in *Thinking: An Invitation to Cognitive Science*, vol. 3, ed. Edward E. Smith and Daniel N. Osherson (Cambridge, MA: MIT Press, 1995). The "strong AI" definition quoted in the text is at p. 417. The script programs Searle targeted are Roger Schank and Robert Abelson, *Scripts, Plans, Goals and Understanding* (Hillsdale, NJ: Lawrence Erlbaum, 1977); the burned-hamburger story and its unspoken answer are Searle's rendering of their question-answering task, and Searle was careful to add that he was not attributing the strong-AI claims to Schank himself (his n. 1) — they belong to the thesis's partisans. Searle names Winograd's SHRDLU and Weizenbaum's ELIZA as equally within the argument's scope (p. 417).

**2.** Searle's response to the systems reply is in his "Minds, Brains, and Programs," 419–420. His key move asks us to imagine the operator internalizing the rules and data banks while preserving the computation. That removes the room's external furniture, but it does not by itself decide the systems advocate's claim that a Chinese-understanding organization may be implemented below the operator's conscious access. The text therefore treats internalization as a sharpening device and answers the stronger reply at the level of the implemented whole.

**3.** The virtual-mind strengthening of the systems reply appears in embryo among the original BBS commentaries' subsystem variants and has been developed most fully by David Cole—"Artificial Intelligence and Personal Identity," *Synthese* 88 (1991): 399–417, with the survey in Cole, "The Chinese Room Argument," *Stanford Encyclopedia of Philosophy*, which separates the virtual-mind version from the simple systems reply and presses it against Searle. The reply in the text is deliberately conditional: naming a virtual subject does not establish the content-integration needed for understanding, but a sufficiently rich implemented organization cannot be excluded merely because its substrate-level operator lacks Chinese.

**4.** Searle's response to the robot reply appears in "Minds, Brains, and Programs," 420. The connection between that reply and the embodiment requirement developed in Part Three is mine, not Searle's; Searle rests on the narrower point that perceptual symbols remain symbols. Varela, Thompson, and Rosch's *The Embodied Mind* (Cambridge, MA: MIT Press, 1991) provides a canonical account of cognition as constitutively shaped by embodied environmental engagement, while Andy Clark's extended-cognition work develops a complementary but distinct account of environmentally distributed cognitive organization. Neither source by itself establishes this chapter's integration test or defeats every robot reply; the argument in the text is the book's.

**5.** John Searle, "Is the Brain a Digital Computer?", *Proceedings and Addresses of the American Philosophical Association* 64, no. 3 (1990): 21–37. The argument treats computational descriptions as observer-relative rather than as intrinsic causal features of brains. This claim has been contested; for a defense see Searle, *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992), ch. 9, especially 207–210 and 225–226, and for a critical response see Daniel C. Dennett, review of *The Rediscovery of the Mind*, by John R. Searle, *Journal of Philosophy* 90, no. 4 (1993): 193–205. Chapter 22 takes up the argument, and the triviality results it connects to, in full.

**6.** Searle, "Minds, Brains, and Programs," 421, where Searle states the combination reply for its Berkeley and Stanford proponents — the robot passage and the "collectively much more convincing and even decisive" billing — and gives the answer quoted here: attribution "rational and indeed irresistible" so long as "we knew nothing more about it," withdrawn once the behavior is known to result from a formal program.

**7.** The distinction between successful artificial performance and understanding is developed explicitly in Luciano Floridi, "AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models," *Philosophy & Technology* 36, no. 1 (2023), article 15. For contrasting positions on language models and meaning, see Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜," *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21), 610–623, especially 617; and Murray Shanahan, "Talking about Large Language Models," *Communications of the ACM* 67, no. 2 (2024): 68–79. These works frame the dispute; the chapter's distinction between mechanism-level content and whole-system understanding is the book's. Fintan Mallory, "Teleosemantics for Neural Word Embeddings," *Mind & Language* (2026), doi:10.1111/mila.70037, secs. 5–5.1, analyzes the training and producer-consumer relation that fixes word-type/context-distribution content. On the account used here, that relation warrants original content where it obtains; neither distal

reference nor whole-model understanding follows. Chapter 23 develops this narrow concession. Shanahan similarly distinguishes the bare model from conversational agents built upon it and argues, on mechanistic grounds, that folk-psychological vocabulary—belief, knowledge, thought—applies only in a qualified, carefully bracketed sense.

**8.** Jason Wei and colleagues defined and popularized the recent machine-learning use of "emergent abilities" in "Emergent Abilities of Large Language Models," *Transactions on Machine Learning Research* (2022): abilities absent in smaller models but present in larger ones and not predictable by simply extrapolating smaller-model performance. In the philosophical taxonomy this is weak emergence — unpredictability from smaller scales — never strong. Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo, "Are Emergent Abilities of Large Language Models a Mirage?," *Advances in Neural Information Processing Systems* 36 (2023), argue more narrowly that, for the model families, tasks, and fixed outputs they studied, nonlinear or discontinuous metrics can create apparent jumps while continuous metrics yield smoother performance curves. The philosophical weak/strong distinction is David J. Chalmers, "Strong and Weak Emergence," in *The Re-Emergence of Emergence*, ed. Philip Clayton and Paul Davies (Oxford: Oxford University Press, 2006); Section 11.5 states the verdict, and Appendix B.7 gives the full treatment.

## Chapter 22: Simulation and the Observer

> *"A map is not the territory it represents, but, if correct, it has a similar structure to the territory, which accounts for its usefulness."*
>
> — Alfred Korzybski, *Science and Sanity*, 5th ed. (1994), 58

CHAPTER OVERVIEW

A bare wall can be matched with countless programs, but a loose match does not make the wall a computer. I grant that a strict chain of causes can make one program, rather than another, a real feature of a physical system in its own right. That result gives us a running structure, not yet meaning, understanding, or a mind. Even a flawless brain model leaves the main question open: it may only copy the powers a brain has, or the concrete system may truly have the right past, links to a world, content, and ties across the whole. We must inspect the running thing, since the words simulation and digital can neither grant those powers nor rule them out.

### 22.1 The Move the Last Chapter Left Open

Chapter 21 closed on a diagnosis: neither operator-level rule-following nor fluent output establishes understanding by the implemented whole. A computationalist can grant all of that and still think the argument has missed him, because (he will say) the mind was never supposed to be the *symbols*. The symbols are just the visible spelling. The mind is the realized computation: recurrence, memory, integration, planning, and control, implemented as one causal organization whether built from neurons or silicon. Attack the operator all you like; the candidate system remains untouched.

That reply forces a distinction the slogan *the program* usually blurs. An abstract program is a formal type, available for indefinitely many implementations and carrying no token history merely by existing. A running implementation is a physical process with determinate causal powers, a provenance, and whatever relations it bears to an environment. This chapter denies mentality to neither by definition. It argues that the abstract type leaves the relevant worldly and system-level facts open, while the running token must be inspected to learn which of them it realizes. A wide computationalist who includes the complete content-bearing and integrative organization can therefore agree with the positive account and disagree only about its name.

> Take the question in its rudest form, the one philosophers play as a parlor game. There is always someone (there is always someone) who announces, with the calm of a person about to enjoy themselves, that the radiator against the far wall is, at this very moment, running Microsoft Word. Not metaphorically. Not "could be wired to." *Running it, now.* The radiator's molecules occupy a physical state too complicated ever to write down, and that state grinds, moment to moment, into

> other states. Write down the sequence of computational states Word passes through as it runs, then pair each one, in order, with one of the radiator's physical states. Given enough physical states to draw on, and enough freedom in how you pair them, the pairing exists. So the radiator runs Word. It also, by the identical trick, runs a chess engine, and the flight software of a 747, and the chess engine's exact opposite, and it dreams of Provence on alternating Tuesdays.

The pairing works on these permissive terms. The question is whether pairing states in a sequence amounts to implementing a program. A proper account must distinguish the radiator from the computer without assuming that only one can think. Causal structure supplies that distinction.

### 22.2 The Wall That Allegedly Runs Any Program

Functionalism defines a mental state by its causal role and permits the same mind to run on different materials. The triviality argument takes that promise and runs it off a cliff. If a mind is the right functional organization, then *having* the mind is *instantiating* the organization, and organization comes cheap. To say a physical system "implements" a finite-state automaton is only to say that some mapping from physical states to automaton states makes the physical transitions line up with the transition table. Take any system with enough internal states changing fast enough (a wall warming unevenly, a bucket of water, the radiator) and, with no rule against gerrymandering, such a mapping always turns up. Putnam proved a sharp form of it: every ordinary open physical system implements every inputless finite-state automaton.[1]

And Putnam inflicted the wound himself: he wanted minds to be organization precisely so they could be multiply realized. The triviality result says the realization relation is *so* unconstrained that implementing-this-automaton achieves nothing at all — a permission slip you write yourself. So the radiator already instantiates the right automaton, and already has my mind, and yours, and the mind of a reader this argument persuades, and the mind of one it leaves cold, all at once, settled only by which mapping a bored interpreter lays over it. A theory on which the radiator already understands everything has not explained understanding. It has mislaid the word.

The point arrives from the other direction too. Searle reached it through physics, not automata. Whether a lump of matter counts as a "symbol," a "computational state," a case of "running a program" is not, he argued, an intrinsic feature of it the way mass and charge are intrinsic. An observer *assigns* that feature to it by reading the computational description on. Syntax, in his flat phrase, is not intrinsic to physics.[2] You can describe the same wall as implementing any program, or its negation, or nothing, depending entirely on the interpretation projected onto its molecular states — and a physical system has the causal properties it has, full stop, whether or not anybody reads them as a computation.

Putnam and Searle, who agreed about almost nothing in the philosophy of mind, reached the same wall from opposite sides — Putnam from inside functionalism, Searle as its standing critic.[3] The natural reply is that the mappings are gerrymandered: genuine implementation must respect causal structure and counterfactuals, not merely trace an arbitrary sequence through matter.[4] That is the strongest repair, and it deserves a clean hearing.

The defect is real and fixable. A repair that respects counterfactuals and the organization of components defeats the wall; a different question survives: whether intrinsic computation fixes meaning.

### 22.3 The Best Repair, Granted and Set Aside

A stronger objection says a careful account of implementation defeats the triviality argument, so the wall does not run every program after all.

Chalmers supplied functionalism's strongest defense two chapters ago; he supplies the implementation repair here too. He grants that *unconstrained* mappings trivialize, then argues that genuine implementation was never unconstrained. For a physical system to implement a computation, its causal structure must *mirror* the computation's formal structure: its states must divide into components that transition under counterfactual-supporting causal laws in an isomorphic pattern. He formalizes this with what he calls combinatorial-state automata. Ma and Kanai press the repair further: the organization must be specifiable without an observer's labels and displayed in the system's intervention profile. An ordinary rock does not thereby implement an arbitrary complex computation: its components must support the specified causal transitions, including those not exercised on a given run.[5] Putnam's mapping laid a story over the system instead of finding one in it. A single system may implement many computations, but choosing a mapping cannot make it implement any computation we please. The wall is innocent.

The repair wins. Intrinsic, counterfactual-supporting causal structure fences the gerrymander out; the radiator, disciplined, does not run Word. Grant all of it.

Chalmers has shown that a system's causal structure constrains *which* computations it implements — a real fact, not a free interpretive gift. The constraint has not shown that fixing the computation fixes a *meaning*. It delivers a pattern in the shape of causal flow. The same abstract transition structure can serve chess, a tax calculation, or nothing anyone cares about. Causal organization fixes the implemented pattern; history and world-relations fix what, if anything, the pattern is about.

Re-instantiation repeats a pattern without deciding whether the later token inherits the first token's content, continues the same subject, understands, or feels. Provenance and world-involving history bear on content; the realized cognitive organization bears on attribution.

### 22.4 What the Wall Lacks

With implementation established, a different question begins: what makes a computational state mean something? A system can really compute while its abstract transition pattern leaves the content open.

Not the formal structure alone. The structure has to come *tethered*: selection, learning, and causal traffic must constrain which worldly conditions a state answers to.[6] A trained producer-consumer mechanism can thereby acquire content through that history. A nervous system may earn *predator* rather than *sandwich crumbs* while leaving finer distinctions unsettled; no interpreter assigned the pairing, but neither did selection carry a taxonomist's pencil. The radiator dreaming of Provence on Tuesdays cannot dream of Provence because it has neither the relevant history nor an

integrated cognitive organization in which such content could figure. A brain has both. Content depends on mechanisms developing and operating in worlds, not on formal organization alone.

The brain's chemistry supplies determinate causal joints, but causal specificity alone is not content; a radiator has determinate chemistry too. Formal computation groups systems by transition structure and omits the selection, learning, and world history that do the tethering.

An artificial system might develop the relevant history too. The implementation repair gives us a genuine computation to investigate; it does not bar its states from acquiring content.

### 22.5 What the Model Does and What the Storm Does

> So the computationalist stops defending the symbols and stops defending the bare notion of implementation, and asks instead for the whole apparatus at once, properly tightened. *Fine*, he says. *Grant me Chalmers's disciplined implementation — counterfactual-supporting causal structure, no gerrymander. Grant me Deutsch's physical principle: the brain is a finitely realizable physical system, so a universal model computing machine can perfectly simulate it. Give me more than a machine with matching inputs and outputs: build a perfect computational duplicate of a living brain (every neuron, every synapse, every ionic flux, tracked and updated in lockstep) and run it. I am no longer claiming the wall thinks. I am claiming that this, the faithful running model of a brain, thinks. What could it possibly lack?*

Grant disciplined implementation, physical simulability, and the perfect duplicate. Which features does the running system merely describe, and which does it actually realize? We have to inspect the deed.

> Imagine a hurricane forecast. The model takes in pressure gradients, sea-surface temperatures, and the rotation of the earth. Its projected storm intensifies, drifts west-northwest, and makes landfall near Galveston. Suppose the forecast proves good enough that Texas evacuates cities on its say-so. Nobody evacuates the data center.[7]

The computationalist can grant all of that: a simulated calculator really adds, and a simulated chess engine really plays chess. Perhaps mentality, too, consists in organization that the running computer actually realizes. If so, the dry data center tells us nothing against its thinking. I agree that the hurricane cannot decide which sort of capacity mind involves.

Watch what the weather model lacks. It updates numerical values for temperature, pressure, and wind velocity through atmospheric equations. The hurricane moves air, lifts heat off a warm ocean, and drives water ashore. The computer's voltages do not produce those atmospheric powers. Refining the grid improves the forecast without turning the calculation into a storm. This

comparison blocks an automatic transfer of powers from model to modeled thing; it supplies no general barrier to artificial realization.

A simulation of metabolism burns no calories — or rather burns only the watts the hardware draws, which is precisely the *wrong* expenditure, the cost of the substrate being itself rather than the cost of a body doing its living. To realize a phenomenon, a system has to possess the **causal powers** of that phenomenon, not merely a description rich enough to predict its behavior.[8]

Coupling can change the answer. A simulated controller wired through transducers to a boiler really controls the boiler. Its variables now participate in the activity previously only modeled. Calculation, control, digestion, and weather have different realization conditions; the word *simulation* tells us none of them.

That shifts the burden to the theory of mind. If consciousness depends only on intrinsic causal-computational organization, a system realizing it is no counterfeit merely because we call it a simulation. If consciousness requires world-involving content, those powers must be realized too. Wire a model into the right traffic and it may acquire a power it previously depicted. The noun settles nothing; what the running system does settles the case.

For more than forty years Searle has pressed the comparison, usually with rainstorms and five-alarm fires: "The brain simulator program by itself no more causes consciousness than the fire simulator program burns the house down or the rain simulator program leaves us all drenched."[9] The hurricane is mine; the analogy is his. Its application to consciousness requires the further argument about what a mind realizes.

### 22.6 Simulability Is Not Realization

The original Church–Turing thesis concerns functions calculable by an effective procedure.[10] Deutsch's physical principle adds a claim about nature: every finitely realizable physical system can be perfectly simulated by a universal model computing machine operating by finite means.[11] His perfect simulation matches the target's input–output behavior under specified labels; his proposed universal machine uses quantum computation. Neither a classical digital machine nor neuron-by-neuron duplication follows just from that principle. The brain duplicate in Section 22.5 therefore adds a stronger stipulation. Physical simulability gives that proposal something it needs, but it supplies no inference from a model's fidelity to mentality. A brain may be perfectly simulable without every power of the brain passing to its model.

### 22.7 The Perfect Brain, and What It Still Lacks

> Give the computationalist the limit case: a simulation tracking every neuron, synapse, neurotransmitter release, and ionic flux in a human brain. Its modeled neurons reproduce the patterns of a living brain having lunch, reading Proust, or falling in love. Hold fixed the fine-grained causal organization inside the brain, including its responses to alternative inputs at the sensory boundary. Leave the new implementation's provenance and environmental relations

> unspecified. Surely, the intuition says, somewhere in there the lights come on.

That stipulation fixes internal organization, not the original subject's entire world-involving history. Two questions remain.

The first gap opens from the side of meaning, the gap Part Three prepared: tracking, selection, and learning relations constrain what a state is about—the causal-historical relations Dretske naturalized and Tye used to secure wide content, which Chapter 6 then identifies with phenomenal character in the conscious case.[12] A match in internal organization does not by itself settle whether a new token inherits those relations. An accidental duplicate has no claim on the donor's coffee-history; a simulation produced by causally copying a particular brain may stand at the end of that history, just as a speaker can inherit a name's reference through a chain she did not begin. Any inheritance would come from the copy's provenance, not from the firing pattern alone. A separately trained copy may instead earn content through its own history. Internal duplication settles neither route. Losing reference to that particular cup would not, however, show that the copy lost the look or taste of coffee. Chapter 6 distinguishes the object referred to from the nonconceptual properties presented in experience; the latter fix phenomenal character. Changed object-reference need not change those properties. A different presented content, no such content, a different experience, and no experience require different arguments.

Ned Block reached the second opening from the functionalist's own side, by way of the homunculus-nation the reader already met in Chapter 20.[13] Here it takes the shape Block gave it in 1995, tailored to precisely this case: the separation of *phenomenal* consciousness (what it is like to undergo an experience) from *access* consciousness, the availability of information for reasoning and control.[14] An exact reproduction of the relevant access functions delivers access consciousness by the functional characterization. Whether it delivers the phenomenal remains the point under dispute, not a conclusion contained in the vocabulary. The perfect duplicate therefore gives the functionalist her strongest case. It does not get to borrow the donor brain's history for free, and the word *simulation* cannot deny it consciousness for free either.

> A determined defender sees the shape of the reply and reaches for the obvious patch. *Granted*, he says, *a brain model on indifferent hardware has no contact with a world. So I will not run a disembodied brain. I will run it inside a simulated body, inside a simulated environment — simulated sensors taking in a simulated world, simulated effectors acting on it, a simulated developmental history. Everything Part Three said a mind needed, my system has. Virtually.*

The patch cannot be dismissed by repeating the word *simulated*. A virtual environment contains real digital objects and real causal relations among the physical states that implement them. A collision can genuinely redirect a process; a virtual predator can genuinely alter what its agent does next. Chalmers is right to insist on that much. The harder question is whether the agent's relations to that environment fix content *for the agent*, or whether the whole ecology remains

a drama whose semantic roles come from its makers. More model is not automatically more grounding. Neither is it automatically none.

The argument does not say no machine or virtual agent could realize a mind. It says that reproducing internal organization does not by itself reproduce the history content may require. A system becomes a candidate understander when content-bearing states participate in a unified, persistent organization of perception, memory, inference, correction, planning, and action. Its environment may contain trees or digital objects; *actual* means causally actual for the system, not pleasingly non-virtual to us.

The concession's best advocate is Chalmers, who has argued at book length that virtual worlds deserve better than their scare quotes: virtual objects are genuine digital objects, virtual events genuinely occur, and a life lived in a well-built virtual world would be a real life in a real environment. He has aimed the point at the present juncture. In his discussion of virtual embodiment, he argues that virtual environments can supply genuine grounding rather than merely stand in for physical ones.[15] He is right. The relations inside a simulated ecology can be causally real, and no general principle makes their identity forever borrowed from a designer. Designers also built the laboratory, named the stimulus, and chose the frog; none of that prevents the frog from earning content in the resulting traffic. The semantic question turns instead on whether selection, learning, and causal history distinguish among equally structure-preserving readings. A bare web cannot (Section 22.8); a mechanism with the right consequential history may. Cups and predators can become genuine contents because the history fixes what the relevant states are for, not because the maker labels the sprites.

Dangerous-for-it marks a separate achievement. A pattern that threatens an organization the system must maintain can ground a practical interest, but it does not tell us what any state is about. Chapter 24 takes up that question separately from meaning and consciousness.

The stronger patch makes the system model *its own continued existence* and act to preserve it — a self-evidencing agent holding its states within viable bounds.[16] Modeling viability is not enough. The regulation has to maintain the organization whose loss would end it; otherwise the self-concern is depicted along with everything else.

### 22.8 The Objection That Takes Mind to Be the Exception

The organizational reply beside the hurricane has a specific semantic version. Vladimír Havlík, writing in *Synthese* in 2025, argues that linguistic understanding has realization conditions a language model can meet.[17] His position, *semantic fragmentism*, grounds an expression's meaning in its relations within a sufficiently rich fragment of language. A corpus reaches what he calls semantic saturation when adding relevant expressions no longer changes those meanings. On his account, acquiring and using that web realizes bounded linguistic understanding. Reference to things outside language need not enter the account at all. A data center need not move air to achieve that.

Havlík and I need not disagree about every bounded achievement. Training can give mechanisms genuine content about linguistic contexts. When several content-guided capacities correct and constrain one another, the book also grants a basis for partial understanding. Neither adult human

breadth nor conscious awareness sets the minimum. But success on a language task, content about its patterns, and understanding through the coordinated use of that content remain distinct claims. A system that uses a correction to revise an inference, a paraphrase, and its later answers gives us more evidence of bounded understanding than one whose successes remain isolated.

A structural limit bears on reference, though it does not refute Havlík's narrower claim. Draw three nodes with an arrow from the first to the second and another from the second to the third, and no other arrows. Read them as three towns joined by direct one-way roads. Now read them as three tanks joined by direct pipes. Stipulate that each domain has exactly those connections. Both readings satisfy the same abstract pattern, although roads and pipes differ. Each node has a distinct position; no symmetry has to swap the nodes within the original graph. What varies here includes both the things assigned to the nodes and the concrete relation assigned to the arrows. Nothing in the bare pattern says *road* rather than *pipe*.[18]

That example deliberately leaves out names, causal history, and evidence that the represented connection really involves roads. Adding those facts may select the first reading. Merely renaming nodes would establish no competing content for a physically trained network. The modest point is that matching a relational form does not, by itself, select its distal interpretation. Newman and Putnam developed related structural and model-theoretic problems; their arguments require attention to which constraints stay fixed. They supply no theorem that a trained mechanism, with its actual history included, must lack determinate content.

Havlík can reply that changing the distal reading while preserving the relevant linguistic relations leaves meaning unchanged in his sense. That reply reaches the argument. A proof that bare structure leaves distal reference open cannot defeat a theory that does not require distal reference. The present result limits what fragmentism establishes about reference; it leaves its narrower understanding claim to further argument and evidence. Our externalist account gives reasons to ask how words connect to the world, but citing that account cannot make Havlík accept its criterion. Nor can I refute bounded linguistic understanding by asking instead for a conscious subject with a human range of concerns.

Rich internal structure still deserves its credit. Researchers have recovered a working map of a game board from a model trained only to predict moves, and "world model" is not an unfair name for what they found. Training made the geometry a network achievement and the mechanism uses it causally. Its history may also fix content, including original content where an interpreting practice does not exhaust the correctness standard. Neither the structure's richness nor the fact that a task came from humans settles that last question. We must examine the history and use rather than rerun the relabeling trick. The familiar merry-go-round objection marks the gap between relations among expressions and distal reference; it cannot decide whether those relations support linguistic understanding.[19]

David Lewis supplied another way to constrain reference: structure *plus eligibility*. Some properties carve the world at its joints, so interpretations assigning those properties stand ahead of gerrymandered rivals.[20] No literal magnet tugs at a symbol. Eligibility constrains the best overall interpretation and may help fix reference. That concession still leaves us to inspect how the

system uses its content. A narrow linguistic achievement may be real without establishing an integrated cognitive life or phenomenal consciousness. The hurricane tells us to investigate the powers actually realized; it cannot deliver the verdict.

**Chapter Summary**

**The claim.** A wall does not run every program because an observer can map its states. A strict causal test can show that a system truly computes, but causal form alone fixes neither meaning nor mind. Calling the system a simulation settles nothing else.

**Where the argument stands.** Putnam exposes how loose mappings trivialize implementation; Chalmers's repair ties computation to the system's own causal structure. Church–Turing results concern computability rather than every modeled power, while the hurricane blocks automatic transfer without deciding the mind case. History and use may fix original content where correctness standards are not exhausted by interpretation. A bare relational pattern leaves distal reference open; that limit does not refute Havlík's narrower claim of linguistic understanding. Content, partial understanding, and consciousness still require separate assessments.

**What's left open.** A real system may combine its own computation, world-directed content, and a unified cognitive life; neither a virtual setting nor an engineered origin bars it. Self-maintenance may add autonomy and interests, but it creates neither meaning nor consciousness.

**The hand-off.** The wall has left the witness stand. Actual systems now require judgment by their histories and organization, not by a loose mapping or the word *simulation*.

**Footnotes**

**1.** Hilary Putnam, *Representation and Reality* (Cambridge, MA: MIT Press, 1988), ch. 5 and the Appendix, esp. pp. 120–125 (the page range at which Chalmers locates the theorem): every ordinary open physical system realizes every abstract **inputless** finite-state automaton. The result requires the system be *open* (interacting with an environment so its states do not cycle) over the relevant interval; the proof constructs the needed state-to-state mapping by exploiting the density of distinct physical states. For automata with specified inputs and outputs, Putnam offers a weaker result: any physical system with the right input–output profile realizes indefinitely many compatible automata, not literally every one. The strength of the inputless claim — *every* automaton, not merely many — is what makes it a *triviality* result rather than a mere restatement of multiple realizability.

**2.** John R. Searle, "Is the Brain a Digital Computer?," *Proceedings and Addresses of the American Philosophical Association* 64, no. 3 (1990): 21–37, at 26–28, where Searle argues that "syntax is not intrinsic to physics" — that computational description is observer-relative, no fact intrinsic to a physical system settling whether it is computing at all. Searle's 1990 paper does work distinct from the Chinese Room of Chapter 21: the Chinese Room argues that syntax does not *suffice for* semantics; "Is the Brain a Digital Computer?" argues that nothing in a physical system is *intrinsically* syntactic in the first place. The restatement in less technical register appears in *The Mystery of Consciousness* (New York: New York Review of Books, 1997), ch. 1. The observer-relativity of syntax is the governing premise; for its extension to semantics see Teodor Negru, "Intentionality

and Background: Searle and Dreyfus against Classical AI Theory," *Filosofia Unisinos* 14, no. 1 (2013): 18–34.

**3.** Putnam, *Representation and Reality*, esp. chs. 5–6. The reversal is genuine and unhedged: the multiple realizability that recommended functionalism in the 1960s is, by Putnam's 1988 lights, exactly what dissolves it once the realization relation is seen to be unconstrained. On the philosophical and rhetorical weight of functionalism's founder arriving at this verdict, see Oron Shagrir, "The Rise and Fall of Computational Functionalism," in *Hilary Putnam*, ed. Yemima Ben-Menahem (Contemporary Philosophy in Focus; Cambridge: Cambridge University Press, 2005). This chapter treats the Putnam–Searle convergence as evidential, not merely rhetorical: independent derivations of one conclusion from incompatible starting points.

**4.** Peter Godfrey-Smith, "Triviality Arguments against Functionalism," *Philosophical Studies* 145, no. 2 (2009): 273–295, esp. secs. 4–5. Godfrey-Smith's inventory separates the versions of the triviality argument that succeed from those that overreach. His response requires realizers to possess the right lower-level causal organization: appropriately similar physical states, independently variable components, and suitably localized role occupants. That retreat from wholly autonomous functional description converges with the position Section 22.4 defends. Godfrey-Smith explicitly brackets teleological versions of functionalism in this paper.

**5.** David J. Chalmers, "Does a Rock Implement Every Finite-State Automaton?," *Synthese* 108, no. 3 (1996): 309–333; the compressed version is in *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press, 1996), ch. 10. Chalmers requires a *combinatorial-state automaton* structure: internal states factor into components whose transitions are reliable, counterfactual-supporting, and isomorphic to the formal computation's state-transition structure. Ron Chrisley, "Why Everything Doesn't Realize Every Computation," *Minds and Machines* 4, no. 4 (1994): 403–420, presses a kindred constraint. For the subsequent debate see Matthias Scheutz, "When Physical Systems Realize Functions…," *Minds and Machines* 9, no. 2 (1999): 161–196; Mark Sprevak, "Three Challenges to Chalmers on Computational Implementation," *Journal of Cognitive Science* 13 (2012): 107–143; and Michael Rescorla, "The Computational Theory of Mind," *Stanford Encyclopedia of Philosophy*. Shuqin Ma and Ryota Kanai push the repair further in "Intrinsic Computational Functionalism: From Observer-Relative Maps to Observer-Independent Structures," arXiv:2606.06424 (2026): their criteria require observer-independent structure grounded in causal-dynamical organization displayed under intervention. The chapter grants disciplined intrinsic realization; it does not claim Chalmers's particular formal account is exceptionless. Sections 7–8 acknowledge logically possible false implementations and multiple computations within one system. The relevant constraint concerns implementation of a specified complex computation, not a small total number of computations per system. Their companion paper, Ryota Kanai and Shuqin Ma, "Intrinsic Computational Functionalism and Simulated Consciousness," arXiv:2606.15348 (2026), rightly shifts the burden: a critic must identify a consciousness-relevant intrinsic structure a purported simulation fails to realize. Chapter 24 accepts that burden without treating thermodynamic self-maintenance as a consciousness requirement; it separates world-involving content, whole-system integration, consciousness-relevant poising and identity, and the autonomy and practical interests a realized stake may ground.

**6.** The teleosemantic grounding gestured at here is developed in Part Three; see Chapter 15 (teleosemantics) and Chapter 16 (semantic externalism), resting on Ruth Garrett Millikan, *Language, Thought, and Other Biological Categories* (Cambridge, MA: MIT Press, 1984), chs. 1–2 on proper functions, and Fred Dretske, *Explaining Behavior: Reasons in a World of Causes* (Cambridge, MA: MIT Press, 1988), for the parallel project through learning history. Artificial training can in principle supply producer-consumer functions and task-relative correctness conditions. Jumbly Grindrod, "Large Language Models and Linguistic Intentionality," *Synthese* 204 (2024): article 71, https://doi.org/10.1007/s11229-024-04723-8, presses the further possibility that corpus-mediated causal-historical chains can support linguistic reference. The text grants both routes while keeping distal reference and whole-system understanding distinct.

**7.** The hypothetical hurricane forecast is mine; the simulation/realization distinction it dramatizes is Searle's (see n. 9). Improving the numerical resolution of a conventional atmospheric model does not give its implementation the atmospheric powers represented. This blocks an inference from predictive fidelity alone to possession of those powers. It does not show that fidelity supplies no evidence about any modeled capacity, or that computation cannot realize mentality. The realization conditions of the particular capacity must be established independently.

**8.** John Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3 (1980): 417–424, at 420–423: the distinction between a system that *has* the relevant causal powers and a system that merely *models* them is central to his reply to the brain-simulator version of strong AI. "Causal powers" is Searle's term of art. The book retains the demand to realize the powers relevant to the attributed capacity without accepting a biological restriction or inferring that a model cannot also be a realization. Sections 22.5–22.7 assess that further question separately.

**9.** John Searle, *The Mystery of Consciousness* (New York: New York Review of Books, 1997), 60. The point recurs across Searle's work from 1980 onward, in a formula he states variously as *simulation is not duplication* and *a simulation should not be confused with the thing simulated.* Searle's rainstorm, fire, and digestion cases are interchangeable illustrations of one structural point: a model of a causal process does not inherit the process's downstream physical effects.

**10.** Gualtiero Piccinini, "Computationalism, the Church–Turing Thesis, and the Church–Turing Fallacy," *Synthese* 154, no. 1 (2007): 97–120, esp. 99–101 and 114–15. Piccinini distinguishes computationalism proper — the empirical thesis that cognition consists in the brain's computation of certain Turing-computable functions — from the far weaker claim that brain functions are Turing-computable. Even the weaker claim requires a physical computability premise; physicalism alone does not supply it. The inference from computability to computationalism is the *Church–Turing fallacy*, and Piccinini's paper shows several historically influential versions of it (Newell, Pylyshyn, and others) to be unsound. The coinage is B. Jack Copeland's — Piccinini writes of "what Jack Copeland has called the Church–Turing fallacy" — see Copeland, "Narrow Versus Wide Mechanism: Including a Re-Examination of Turing's Views on the Mind-Machine Issue," *Journal of Philosophy* 97, no. 1 (2000): 5–32. The thesis and Deutsch's physical principle are distinct claims doing distinct work — and the computationalist tends to draw on whichever one is not currently under examination.

**11.** David Deutsch, "Quantum Theory, the Church–Turing Principle and the Universal Quantum Computer," *Proceedings of the Royal Society of London A* 400 (1985): 97–117, at 99. Deutsch's physical principle says that "every finitely realizible physical system can be perfectly simulated by a universal model computing machine operating by finite means." It is an additional thesis about physics, not a consequence of physicalism. The crucial point is that the principle concerns *simulation*: it is silent on whether running the model instantiates every causal power of the system modeled. Deutsch defines perfect simulation as computational equivalence under specified input/output labels, not fine-grained organizational duplication; the latter is a separate stipulation here. His universal model is quantum, not simply a classical digital computer. The principle remains contested in its strongest form; granting it supplies physical simulability without settling whether a running model realizes mentality. See sec. 1, pp. 3–4 of the author's later typeset version, https://www.daviddeutsch.org.uk/wp-content/deutsch85.pdf, for the definition and quantum qualification.

**12.** Michael Tye, *Ten Problems of Consciousness* (Cambridge, MA: MIT Press, 1995), develops the PANIC theory on which phenomenal character is identical to representational content that is Poised, Abstract, Non-conceptual, and Intentional; Fred Dretske, *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995), and earlier *Explaining Behavior: Reasons in a World of Causes* (Cambridge, MA: MIT Press, 1988), ground content in tracking relations secured by the system's design or learning history. Tye's abstract-content requirement excludes particular concrete objects from phenomenal content (with limited exceptions for the subject); see *Ten Problems*, pp. 137–39. Chapter 6 likewise distinguishes singular reference and accuracy from the nonconceptual property content that fixes character. Its strong representationalist identity concerns content of the appropriate kind under the relevant conscious organization, not every difference in reference. World-involving history bears on which properties a state represents; poising and the book's further account bear on whether the state is conscious. Failure to inherit reference to the donor's particular object proves neither different phenomenal content nor absent experience. A merely formal description does not establish these conditions, but a running artificial system may realize them.

**13.** The homunculus-nation and the absent-qualia objection are built and sourced in Chapter 20, Section 20.4.1 (notes 13–18); the primary text is Ned Block, "Troubles with Functionalism" (1978). The present chapter draws only on the later cut tailored to the brain-simulation case — Block's access/phenomenal distinction (n. 14).

**14.** Ned Block, "On a Confusion about a Function of Consciousness," *Behavioral and Brain Sciences* 18, no. 2 (1995): 227–247, esp. 227–30. The phenomenal/access distinction has organized the consciousness literature since: access consciousness is functionally definable (information poised for reasoning, report, and the control of action); phenomenal consciousness is what it is like to undergo a state. Reproduction of the stipulated access functions secures the first under that characterization; it does not, by the distinction alone, settle the second. Block's taxonomy is not a proof that functional duplicates lack phenomenal consciousness.

**15.** David J. Chalmers, "Could a Large Language Model Be Conscious?," *Boston Review*, August 9, 2023 (from a NeurIPS 2022 address); see the discussion of virtual embodiment. The book-length

defense of virtual realism is *Reality+: Virtual Worlds and the Problems of Philosophy* (New York: W. W. Norton, 2022): virtual objects are genuine digital objects, virtual events genuinely occur, and virtual experience need involve no illusion. The text grants the metaphysics. Designer-assigned labels do not by themselves fix what an agent's states represent, but selection, learning, and consequential traffic within a virtual ecology can in principle do so. Virtuality neither supplies grounding nor blocks it; self-maintenance bears separately on autonomy and interests.

**16.** The "self-evidencing" idiom belongs to Jakob Hohwy, "The Self-Evidencing Brain," *Noûs* 50, no. 2 (2016): 259–285: a system that embodies a model of itself gathers evidence for its own existence by acting to keep its sensory states within the bounds the model expects. The underlying formal framework is Karl Friston's free-energy principle — organisms resist disorder by minimizing variational free energy, an information-theoretic bound on the surprisal of their sensory states: Karl Friston, "The Free-Energy Principle: A Unified Brain Theory?," *Nature Reviews Neuroscience* 11 (2010): 127–138. Neither author draws the present distinction. The question is not whether self-evidencing can be implemented in software in the abstract, but whether a particular physical implementation's regulation helps constitute the maintenance of the very organization doing the regulating. Chapter 24 calls that condition thermodynamic self-maintenance and treats it as substrate-general.

**17.** Vladimír Havlík, "Meaning and Understanding in Large Language Models," *Synthese* 205, art. 9 (2025), https://doi.org/10.1007/s11229-024-04878-4. See the discussions of referential grounding, semantic fragmentism, and the grounding of LLM meanings. Havlík's *semantic fragmentism* holds that LLMs achieve genuine, bounded linguistic understanding because training embeds their internal states in dense contextual and inferential relations, supports minimal semantic contents, and can ground meanings without requiring conscious awareness or direct sensory contact. Havlík explicitly rejects referential grounding as necessary for linguistic meaning; on his view an expression represents its place in the corpus's relational complexity. A limit on distal reference therefore does not refute semantic fragmentism's narrower claim. The chapter preserves that distinction and leaves the sufficiency of his linguistic-understanding criterion open, rather than demanding conscious subjecthood or adult human breadth as an unargued additional condition.

**18.** For related limits on pure structural description, see M. H. A. Newman, "Mr Russell's 'Causal Theory of Perception,'" *Mind* 37, no. 146 (1928): 137–148, and Hilary Putnam, *Reason, Truth and History* (Cambridge: Cambridge University Press, 1981), ch. 2 and the Appendix. These are related but distinct arguments, not one theorem that directly applies to every trained network. Putnam's ch. 2 considers reinterpretations preserving whole-sentence truth values while predicate extensions change. The main-text example illustrates transfer of a relational structure between domains, including transfer of the relation's interpretation; it does not infer a nontrivial automorphism of a fixed graph from the existence of an isomorphic copy. Actual names, training relations, and causal history are excluded from that toy example, not shown semantically idle. Selection history and corpus-mediated causal chains may fix content or distal reference. Chapter 23, note 8, makes the same limited distinction. For rich, behavior-controlling internal structure recovered by interventionist probing, see Kenneth Li et al., "Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task," ICLR 2023 (arXiv:2210.13382). Such

evidence concerns realized internal organization and its use, not a proof of semantic independence or whole-system understanding. Where an interpreting practice exhausts the correctness standard, content remains derived; inheritance of a human-designed task alone proves no such exhaustion.

**19.** The "merry-go-round" is Stevan Harnad's image for the *symbol grounding problem*: symbols defined only by other symbols, circling endlessly without ever touching what they are about. Chapter 23 (Section 23.2) gives the problem its source and its bearing on language models. The point here is limited to distal reference: a web of intra-symbolic relations does not *by itself* select a worldly interpretation, though training history and corpus-mediated causal chains may add what bare geometry lacks. This does not establish that corpus-relational meaning or bounded linguistic understanding is impossible.

**20.** The reply is David Lewis's, "Putnam's Paradox," *Australasian Journal of Philosophy* 62, no. 3 (1984): 221–236, esp. 227–29, which answers the permutation argument by adding a constraint of *eligibility*—reference biased toward the more natural properties—to the structural ones Putnam allows. J. R. G. Williams reconstructs and presses the eligibility view in "Eligibility and Inscrutability," *Philosophical Review* 116, no. 3 (2007): 361–399, esp. 372–80. Eligibility constrains interpretations globally; it need not operate through a causal "pull" on a particular symbol. The text grants that naturalness can break some structural ties and may help fix reference when joined to selection history and causal engagement. That semantic concession leaves whole-system understanding as a separate attribution question. For sustained skepticism about treating reference magnetism as a Lewisian doctrine at all, cf. Wolfgang Schwarz, "Against Magnetism," *Australasian Journal of Philosophy* 92, no. 1 (2014): 17–36.

## Chapter 23: Form Without World

> *"O, a hyper-intelligent deep-sea octopus who is unable to visit or observe the two islands, discovers a way to tap into the underwater cable and listen in on A and B's conversations."*
>
> — Emily M. Bender & Alexander Koller, "Climbing towards NLU"

---

CHAPTER OVERVIEW

Imagine an octopus that hears every word sent between two islands yet never meets the world the speakers discuss. That case shows why form alone cannot fix what words point to. Training can still give a mechanism original content about word distributions, conditional on the history and use Mallory describes. Text, cameras, and tools may also connect its states to the wider world. I distinguish those semantic gains from understanding: the use of content across capacities that can constrain and correct one another. That use may support understanding within one task or encounter. Enduring agency requires more, and neither achievement settles consciousness.

### 23.1 From the Limit Case to the Live One

Language models explain, summarize, translate, write usable code, hold a thread across a long conversation, and adjust their register to yours. I use these systems daily, and dismissing that as "autocomplete" refuses to look. Their builders did not simulate a cortex. Training them to predict text produced capacities whose extent and organization now need explaining.

Chapter 22 showed why the label *simulation* cannot settle whether a running system realizes a mind. The chat window poses a more specific question: what can a history of learning from text establish? Three things need separating. Linguistic form supplies patterns among signs. An actual corpus also has a history connecting those signs to speakers and their world. The running system may then use what training established in ways that support further learning and correction. A limit on form alone need not settle what those wider relations achieve. One more clarification: the world in this chapter's title is the world words are about. The actual occurrences of words belong to the world too, and Section 23.3 grants that training on them can give content at that grain. What form alone cannot supply is reference to the trees, tides, and people the words describe.

### 23.2 The Eavesdropping Octopus

Two computational linguists give the cleanest statement of the gap. Emily Bender and Alexander Koller built their argument to discipline their own field: to puncture the slide, common in the research literature, from *the model produces fluent text* to *the model understands*. They made it with a thought experiment.[1] Here is an adaptation of their story; the narrower conclusion drawn from it is mine.

> Two people, A and B, live on separate islands and talk to each other through an undersea telegraph cable. A hyper-intelligent octopus, O,

> taps the cable and listens. It never sees the islands, never meets A or B, never observes a single thing the two of them talk about — it has access only to the *form* of the signals, the patterns of symbols flowing back and forth. And O proves an extraordinary student of those patterns. It learns what tends to follow what so well that one day it cuts B out of the loop, impersonates him, and keeps up B's end of routine social conversation. A notices nothing. Across the wire, within that task, O performs brilliantly.
>
> Then a bear appears on A's island. A, panicking, grabs a couple of sticks and begs B for instructions to turn them into a weapon. O may emit a useful-looking answer, perhaps even a lucky one. Familiar conversation alone does not establish the worldly basis on which B's answer would answer to this bear, these sticks, their material possibilities, and the consequences for A. O may have learned real relations within the conversation. Its grasp of formal patterns alone does not establish what "stick" or "bear" refers to. The actual conversation's history may transmit reference; that still does not establish practical understanding of these sticks and this bear. The crisis exposes not a guaranteed behavioral failure but a missing basis for success, a gap that routine conversation could conceal. (B would have known which end of the stick to sharpen.)

Bender and Koller predict that O will fail when the bear arrives. I retain a narrower conclusion: success at familiar conversation does not by itself explain a capacity to deal with this unfamiliar worldly problem. That gives us the form/world gap without settling consciousness. Bender and Koller do not claim O lacks an inner life; they claim O lacks direct contact with the world its words concern. Distal reference and communicative understanding depend on a relation between form and what a speaker uses it to *do*, grounded outside the text in a shared world and communicative intent.[2] The octopus restages one lesson of the Chinese Room while changing the diagnosis: Searle stresses syntax without intrinsic intentionality; Bender and Koller expose form without a shared world or communicative intention. The contemporary language model runs a far larger and less isolated version of the feat. Training history and feedback complicate the comparison.

Behind the octopus sits Harnad's *symbol grounding problem*: definitions made only of more symbols never supply a worldly address.[3] A model's word-types descend from human practices, while training shapes mechanisms for predicting relations among them. The first route may transmit public reference; the second can establish original content about linguistic distributions under the conditions described below. Grounding, Harnad stressed, is not a full theory of understanding.[4]

### 23.3 Distributional Semantics: Structure Without an Address

The octopus and the merry-go-round say what a form-only system lacks; they do not say what it has. Behind every language model lies a serious wager about meaning: *distributional semantics*. The case against these systems holds only as far as it meets that wager at its strongest.

Its founding idea fits in a sentence the field still recites. In 1957 the British linguist J. R. Firth wrote, "You shall know a word by the company it keeps."[5] A word's meaning shows itself in the words it tends to appear beside. *Dog* runs with *bark*, *leash*, *vet*; *cat* with *purr*, *whiskers*, *litter*; and the company the two share (*pet*, *fur*, *paw*) already marks what their meanings hold in common. Zellig Harris, the American structuralist, had given the point its stronger, testable form three years earlier, in what we now call the *distributional hypothesis*: words with similar distributions have similar meanings. A word's profile of contexts can therefore supply evidence about its meaning. The stronger claim that this profile exhausts meaning needs a further argument. Distributional prediction underwrites text pretraining in the models discussed here; multimodal learning and later tuning add other objectives.

Machine learning turned the idea into arithmetic. Word vectors place terms with similar company near one another, yielding clusters and relations that permit the famous *king* − *man* + *woman* ≈ *queen* result.[6] Transformers go further, constructing context-sensitive representations across layers, so *bank* beside *river* differs from *bank* beside *loan*. The machinery has moved far beyond word2vec, but remains distributional because context trains the representation.

Concede one thing more. Training does not merely deposit a chart for a later reader; it selects mechanisms for using one stretch of linguistic context to predict another. The resulting states acquire determinate causal roles, learned sensitivities to textual distributions, and effects on later processing. Fintan Mallory's consumer-based teleosemantics argues that this history can give word embeddings genuine, original content about linguistic contexts.[7] His claim reaches beyond geometry: a learned mapping operates inside the mechanism, while errors against actual occurrences shape both its producer and its consumer.

The target must stay in view. Suppose a model predicts *engine* beside *cat*, although *pet* dominates those contexts in the sampled text. Human beings supplied the words; they cannot change those occurrences by approving the answer. A corpus can constrain accuracy independently of a judge's verdict even though human practice constituted its word-types. This is a claim about contexts, not about cats.

The matched test introduced in Section 14.2.2 helps distinguish targets. Give both learners selection, error correction, and consumers that use their embeddings. Let one history train against actual sampled contexts and the other against an approved-answer schedule that usually matches them. Separate the schedule from the samples while keeping the word key and context window fixed. Then vary the samples while leaving the schedule fixed. Trace which discrepancy retunes the embedding and changes the consumer's prediction, even if a score bonus masks the failure. These histories can support different contents despite matching ordinary answers. A schedule-tracker may have original content about the schedule; it has not thereby earned our further reading about actual contexts. The comparison identifies what a state represents, not whether linguistic content must always be derived.

Mallory already describes the relevant dependency: output is compared with actual co-occurrences, and the discrepancy changes the mechanisms that produce and consume the embedding. On the teleosemantic account defended here, that history warrants original content about word-

types and their context distributions, conditional on the algorithm operating as described. No additional experiment must supply a missing semantic ingredient. A particular implementation still needs evidence that its training and use instantiate that relation; a published algorithm does not certify every model. This concession concerns occurrences of words in data. It establishes neither reference to a particular cat nor understanding by the containing system.

Coordinate a change of indices, matrices, and word key, and the same word-to-context mapping can survive. Change which actual word feeds an index, or redefine the context window, and a candidate content-fixing relation changes. Those interventions cannot be counted as mere changes of interpretation. Nor would switching off learning erase an established history: ongoing correction supplies evidence here, not a condition every representation must satisfy.

Distribution captures structure: relations among states and the transitions the mechanism expects. It does not by itself fasten that pattern to a distal referent. Hold the corpus's formal pattern fixed while varying its proposed links to trees or animals; geometry alone cannot choose. Actual causal history may constrain those links. Content about a word's distribution, even if original, does not by itself fix which particular tree a use concerns.

The complaint predates machine learning. In 1928 M. H. A. Newman observed that the structure of a domain leaves its relata open: any suitably arranged collection of the right size can satisfy the same pattern. Putnam later pressed the same point in his model-theoretic argument. A distributional space inherits the result. Its selected function constrains useful interpretations; its geometry cannot say which word lands on which worldly thing.

Iwan Williams presses exactly on the phrase *by itself*. Bare structure leaves its worldly address open; a selected learning history may privilege one correspondence over its rivals and help fix which worldly structure a model's states answer to. Whether current models meet those conditions remains an empirical question, not something Newman's theorem settles from the armchair.[8]

Grant the possibility. A privileged correspondence may fix what a trained state represents, including original mechanism content where a world-involving history supplies a standard independent of interpreting practice. That still leaves the level of attribution open. Do those states constrain and correct one another within a task? Does that integrated use continue across encounters? The first asks about understanding; the second about an enduring agent. Distributional semantics gives us more than an inert shadow: content can become causally active machinery while the containing system's understanding and consciousness remain to be shown.

That, finally, gives the next objection its shape. If meaning lives in the geometry of a vector space, hope naturally turns to grounding the *vectors* themselves, tying them to the world by something other than more text. That hope has lately gathered real force.

### 23.4 The Multimodal Rejoinder

> Here the careful reader raises a hand. *Your premise is already obsolete*, the objection runs. *The octopus listens only to a cable; it is trapped in pure form. But the systems we actually use now are not text-only. They*

> *take images, so they "see." They call tools and take actions in software environments, so they "act." They are tuned by reinforcement learning from human feedback, so their behavior is shaped by people responding to what they do. Pixels, sensor streams, actions, reward — surely these break the confinement to form. Surely the modern model has reached past the cable and touched the ground.* This is the robot reply of Chapter 21, returned in modern dress.

That objection has recently received its strongest statement. Dimitri Coelho Mollo and Raphaël Millière have made probably the best contemporary case that large language models can escape the grounding problem, and they make it the way the debate has badly needed, trading slogans for distinctions.[9] Their first move cleans up the question. "Grounding," they point out, names at least five different things in the literature, run together to everyone's confusion; the one that bears on whether a system's states reach extra-linguistic reality is what they call *referential* grounding: a representation's standing in the right relation to its worldly referent, independent of any meaning a human reads into it. Sensorimotor transduction, they argue, supplies only one possible route. Causal-informational relations to the world, combined with a training history that selects states for responding to it, may privilege a genuinely referential interpretation. The octopus, on this view, told a story about a system frozen in a regime that current systems may have left behind.

Sensorimotor transduction gives one route to reference, not the only one. A live camera supplies causal contact; joined to selection or learning, it may help fix content. Contact alone supplies neither cognitive integration nor consciousness.

> By *the system* I mean the deployed organization proposed as the understander, including memory, sensors, and tools when they form parts of its causal loop — not the surrounding company merely because it trains, powers, or repairs the candidate. The boundary must match the capacity claimed: one episode may support task understanding, while learning carried across encounters calls for a persisting organization. Autonomy and welfare require a separate test.

Understanding something, on the account used here, requires more than a content-bearing state: the system must use that content to connect inferences, expose conflicts, and guide correction within the domain. Those capacities need to constrain one another. Otherwise a successful answer may coexist with an unrelated guess and an unused correction, leaving no explanation of why the system grasps the relation among them. This integrated use is the commitment. Memory, planning, flexible action, and retained learning supply defeasible evidence for its breadth and durability, not a checklist every understander must complete.

Consider a short-lived solver that reasons through an algebra problem, checks the result by substitution, and changes its next inference when the check fails. Suppose the same representation explains those responses when the problem changes, rather than merely cueing a memorized answer. That organization could support partial understanding within the episode even if all

memory vanishes afterward. Forgetting does not retrospectively undo the grasp. It prevents that episode alone from establishing an enduring agent. Whether an actual model has the proposed organization remains an evidential question.

Human feedback and automated verifiers sharpen the point. A reward signal really does sort answers; a compiler really can return success or failure.[10] Feedback from outside presents no defect: children, animals, and any plausible artificial learner depend on external causes. In current model training, correction can select mechanisms whose states reliably discriminate features relevant to a platform-chosen task. Whether those states possess original content depends on what gives the task and its verdicts semantic significance. Whether the deployed model understands depends on how it integrates them.

Consider a model trained to produce code that passes a checker. The intended task might concern correct arithmetic; the checker might test only whether the program compiles. Those standards differ before anything goes wrong. Now introduce a fault that makes the checker accept a bad program. Reinforcement follows the verdict. That identifies the immediate feedback channel, but it does not establish that every relevant state represents approval. A state historically recruited to track program correctness can now misrepresent under that established standard. Human approval and accuracy likewise come apart in the documented sycophancy results.[10a]

Chapter 15′s comparison asks more than which wire carries reward. Did the relevant state develop through tracking compilation, arithmetic correctness, or approval? Which later consumers use it, and what happens when those properties diverge? Hold the checker's verdict fixed while an independent execution test reveals the arithmetic error. If that discrepancy corrects the shared state and changes later code, we gain evidence about what the state answers to. If only the reward channel governs revision, that limits the evidence for a wider standard; it does not erase a previously established one. The same test applies to a child with an unreliable teacher. Corrupt teaching can teach an error without instantly changing the subject matter.

Interpretability and in-context learning results supply further evidence about internal organization, not a shortcut past this historical question.[10b] Training can internalize a borrowed standard. It may also establish world-answerable functions when the relevant history and use support them. Neither a human grader nor an engineered checker decides between those possibilities merely by supplying the feedback. Original content and whole-model understanding remain separate claims.

Give Mollo and Millière's grounding story its strongest causal reading: causal-informational relations plus selection history may privilege one world-involving interpretation of model states over others. That concession grants genuine referential content where the conditions are met.[11] The literature leaves the reach and grain of model reference contested, and the strongest positive case deserves its authors' names. Matthew Mandelkern and Tal Linzen argue that a model's words may refer the way many of ours do: through the causal-historical chains of the linguistic community whose text trained it—Kripke's picture, run through a corpus.[12] Grant the chain; Chapter 12 already showed how testimony, deference, and causal history can transmit a public address to consumers who never meet the referent. No ownership supplement is needed to make

the reference real. What remains to be shown is how the candidate uses those contents across its capacities, and how far that integrated use extends.

Components do not settle the question by accumulation. Cameras, memory, tools, verifiers, and online correction supply evidence of understanding when their contents constrain and correct one another within the candidate organization. Chapter 24 asks how far those competences compose a unified cognitive life.

### 23.5 Why "Hallucination" Misnames It

One borrowed word can quietly damage clear thinking about these systems. When a model produces a footnote citing a paper that does not exist (the author never wrote it, the journal volume runs ten issues short, the page numbers point into empty air) engineers say the model *hallucinated*. As operational shorthand for a familiar failure class, the term does no harm. Trouble begins when the shorthand starts doing philosophy.

They found a word that already had a job, and the job mattered. In philosophy, *hallucination* names something specific: a subject undergoes an experience that *seems*, from the inside, to present a worldly object (a pink rat on the counter, a friend at the foot of the bed) when no such object is there. The experience occurs; the object does not; and the whole philosophical delicacy lives in the *seeming*. Tim Crane, working from an intentionalist account of the kind this book defended in Part Two, treats hallucination as a state whose content fails to match how things stand.[13] Mike Martin, defending disjunctivism, denies even that much shared structure between perceiving and hallucinating.[14] Take whichever side you like; both require a *subject who seems to see something*. A hallucination, on any account worth having, happens to someone.

Now look at what the engineer names. A system trained on text emits a string that mentions a paper that does not exist. An internal mechanism may genuinely misrepresent its selected target, and the output is false under the public standards governing the words. Those are real semantic failures. What does not follow is the phenomenal claim folded into *hallucination*: that something surveys a citation field and seems to find a paper there. Misrepresentation need not be experience. On the chapter's argument, *hallucination* still names the wrong shape when treated as more than engineering shorthand.

Teleosemantic misrepresentation belongs first to content-bearing mechanisms: a producer can fail to indicate the condition its consumer was shaped to use.[15] Public language adds another level on which output can be false. Neither error requires an owner. Calling the model a subject that *hallucinates* requires further evidence that these contents integrate into a point of view.

### 23.6 What the Negative Case Clears the Way For

The negative conclusion established here is limited: linguistic form alone cannot fix distal reference, and acquiring content does not by itself establish whole-system understanding. A bare model, a conversation with temporary task memory, and a persistent tool-using agent present different candidates. None receives a verdict merely by removing or adding a component. This chapter has not conducted a model-by-model assessment of their integration; Chapter 24 takes up the criteria and the relevant evidence. Partial understanding within an episode, enduring agency, and phenomenal consciousness remain distinct questions.

**Chapter Summary**

**The claim.** Language models learn linguistic form without modeling a brain. Training can make content causally indispensable, including original content at a specified grain. Form alone cannot fix distal reference. Content in a mechanism does not by itself make the whole model an understander.

**Where the argument stands.** Mallory's described training relation warrants original content about word distributions on this book's teleosemantic account; particular implementations require their own evidence. Corpus-based chains may also preserve public reference. Understanding requires integrated use of content across capacities that constrain and correct one another. That use may support a partial, domain-specific grasp within one episode without establishing an enduring agent. Neither semantic content nor such a grasp establishes consciousness.

**What's left open.** The reach and grain of model reference remain disputed, as does the boundary of the candidate system. The extent of understanding in actual models requires assessment of their organization. Subjecthood, consciousness, autonomy, and welfare call for separate judgments.

**The hand-off.** The black jigsaw now has windows and hands. The next chapter asks when those parts form one system rather than a useful pile.

**Footnotes**

**1.** Emily M. Bender and Alexander Koller, "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), 5185–5198; the octopus appears at 5188–5189. Bender and Koller frame the argument around a distinction between *form* (the observable marks or sounds of language) and *meaning* (the relation between forms and communicative intent), and argue that "the language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning" (5185). Their target is the same overclaiming this Part resists — the inference from fluency to understanding — reached from computational linguistics rather than philosophy of mind. For a critical reply defending the sufficiency of form-derived structure, see Julian Michael, "To Dissect an Octopus: Making Sense of the Form/Meaning Debate," online essay, 2020, julianmichael.org.

**2.** The reliance on a *shared world* and *communicative intent* aligns the octopus with the externalist and Gricean traditions of Chapter 16 rather than with Searle's biological naturalism, and the two diagnoses, though they reach one verdict, differ in emphasis: Searle locates the Chinese Room's failure in the absence of intrinsic intentionality, while Bender and Koller locate the octopus's failure in the absence of a grounding relation between form and the speakers' shared world. This book takes the example to isolate what form alone leaves undetermined, not to establish the absence of causal-communicative relations from every actual model's history. Corpus-mediated reference and later corrective use receive separate treatment in Section 23.4.

**3.** Stevan Harnad, "The Symbol Grounding Problem," *Physica D* 42 (1990): 335–346, esp. 338–45. Harnad's diagnosis: symbol meaning cannot be intrinsic to a symbol system, because within the system there is nothing but more symbols; his proposed remedy is *sensorimotor grounding*, an internal mechanism connecting at least some elementary symbols to the categories of things the

system can pick out in perceptual interaction with the world, so that higher symbols built from them inherit a connection to the world. The problem is the precise, contemporary form of the worry the Chinese Room dramatizes.

**4.** Harnad is careful to distinguish symbol grounding from a full theory of meaning: connecting a symbol to a worldly category does not by itself establish understanding or inner life. The book keeps that distinction. Grounding *distal reference* to an extra-linguistic particular differs from fixing content about linguistic distributions, which may itself involve a worldly relation (note 7). Steven T. Piantadosi and Felix Hill, "Meaning without Reference in Large Language Models" (2022), arXiv:2208.02957, defend genuine aspects of meaning without reference or sensory grounding. The position has a distinguished pedigree in conceptual- and inferential-role semantics, from Wilfrid Sellars through Robert Brandom, *Making It Explicit* (Cambridge, MA: Harvard University Press, 1994), and Ned Block's functional-role account. Block's two-factor theory, "Advertisement for a Semantics for Psychology," *Midwest Studies in Philosophy* 10 (1986): 615–678, pairs internal conceptual role with a world-involving factor because role alone yields narrow content without distal reference. Mollo and Millière likewise allow inferential-role grounding where a vocabulary's role is exhausted by inference. Anders Søgaard, "Understanding Models Understanding Language," *Synthese* 200, no. 6 (2022): article 443, defends both inferential and referential semantics for transformer models. Brandom's normative pragmatics remains a powerful account of public meaning and conceptual practice. None of these semantic concessions decides whether a text-only model integrates content into whole-system understanding.

**5.** The slogan is J. R. Firth's, from "A Synopsis of Linguistic Theory, 1930–1955," in *Studies in Linguistic Analysis* (Oxford: Blackwell, 1957), 1–32 ("You shall know a word by the company it keeps," at 11). Its precise form is Zellig S. Harris's *distributional hypothesis* — "difference of meaning correlates with difference of distribution" — in "Distributional Structure," *Word* 10, nos. 2–3 (1954): 146–162. For the modern theory grown from it, see Alessandro Lenci, "Distributional Models of Word Meaning," *Annual Review of Linguistics* 4 (2018): 151–171, and Gemma Boleda, "Distributional Semantics and Linguistic Theory," *Annual Review of Linguistics* 6 (2020): 213–234. Jumbly Grindrod (cited below) reads the success of large language models as a "partial vindication" of exactly this approach — a reading this section grants, then qualifies.

**6.** The geometric form of the hypothesis is the *word embedding*: Tomáš Mikolov et al., "Efficient Estimation of Word Representations in Vector Space" (2013), arXiv:1301.3781 (the word2vec models), and Jeffrey Pennington, Richard Socher, and Christopher D. Manning, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (2014), 1532–1543. The *king − man + woman ≈ queen* regularity is reported in Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig, "Linguistic Regularities in Continuous Space Word Representations," in *Proceedings of NAACL-HLT 2013*, 746–751. Modern transformers replace the one-vector-per-word picture with contextual token representations generated across layers by attention; see Ashish Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems* 30 (2017), and Jumbly Grindrod, "Large Language Models and Linguistic Intentionality," *Synthese* 204 (2024): article 71. The distributional lineage remains, but the architecture does not merely scale word2vec by adding parameters.

**7.** Fintan Mallory, "Teleosemantics for Neural Word Embeddings," *Mind & Language* (online 2026), esp. secs. 3–5.1, https://doi.org/10.1111/mila.70037, argues that word2vec-style embeddings acquire genuine, original rather than derived content through a consumer-based teleosemantics. In the two-matrix model, M produces the embeddings, M′ consumes them to generate outputs, and backpropagation compares those outputs with actual context distributions and modifies both matrices. Sections 4 and 5.1 (author manuscript, pp. 7–8 and 11–13) describe the historical-mechanistic relation that warrants this chapter's conditional concession. It is conditional on that relation obtaining and on the teleosemantic account, not a theory-neutral proof or a reproduced experiment. No new matched intervention is required merely because the article does not report one. The comparison borrowed from Chapter 14 distinguishes context-directed from schedule-directed content; either may be original at its own grain. A particular implementation's selection history and consumer use remain matters for verification. Mallory's target concerns actual word-types and contexts, not a word's extra-linguistic referent. He distinguishes learned embeddings from keyed input encoding; coordinated reparameterization preserves the mapping, while changing the external word key changes a potentially content-fixing relation. His sec. 5.3 also leaves transferred embeddings and their new consumers for further work. Jumbly Grindrod, "Large Language Models and Linguistic Intentionality," *Synthese* 204 (2024): article 71, https://doi.org/10.1007/s11229-024-04723-8, develops a distinct route through public practice and corpus-mediated causal history. Neither route establishes whole-model understanding. Chapters 14 and 15 reach the same conditional verdict; this chapter applies it to the form-and-meaning debate.

**8.** That a system of relations fixes its relata only up to structure—leaving distal reference undetermined by structure alone—is an old result, not a new complaint: it is the burden of M. H. A. Newman's 1928 objection to Russell's structuralism ("Mr Russell's 'Causal Theory of Perception,'" *Mind* 37, no. 146 [1928]: 137–148) and of Hilary Putnam's model-theoretic argument (*Reason, Truth and History*, Cambridge: Cambridge University Press, 1981, ch. 2 and the Appendix). A distributional space inherits that limited result: its geometry alone cannot fix which worldly thing a word is about. Iwan Williams, "Can Structural Correspondences Ground Real-World Representational Content in Large Language Models?," *Mind & Language* (2026): 1–19, doi:10.1111/mila.70018, argues that a correspondence can ground real-world content when the selected task privileges it and alternatives are eliminated. The text grants that a learning history may establish original mechanism content where the privileged correspondence fixes world-answerable correctness not exhausted by an interpreting practice. Causal privilege alone does not settle that independence, but a distributional target does not disqualify it.

**9.** Dimitri Coelho Mollo and Raphaël Millière, "The Vector Grounding Problem," *Philosophy and the Mind Sciences* 7, no. 1 (2026): 1–28, esp. secs. 4–6.2; preprint arXiv:2304.01481 (2023). Mollo and Millière distinguish five notions of grounding—sensorimotor, epistemic, communicative, relational, and referential—and argue that only *referential* grounding is necessary for a representation to be about extra-linguistic reality. Their positive case combines causal-informational relations with a feedback-driven selection history, especially RLHF. The chapter grants that these conditions can help fix extra-linguistic reference. Section 6.2.1 of the revised preprint (arXiv:2304.01481v3, 2025) expressly distinguishes the designed function of the learning apparatus from the selected functions of its internal states, and intended rewards from the properties

feedback actually and reliably tracks. The chapter's criterion requires world-answerability not exhausted by an interpreting practice; engineered or social inheritance alone does not establish failure of that criterion. The further attribution of understanding to the whole model remains separate.

**10.** Reinforcement learning from verifiable rewards (RLVR) optimizes a model against programmatically checkable outcomes—unit tests, proof checkers, exact-match answers—rather than human preference judgments. The term and recipe enter with Nathan Lambert et al., "Tulu 3: Pushing Frontiers in Open Language Model Post-Training," arXiv:2411.15124 (2024); a prominent demonstration appears in D. Guo et al., "DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning," *Nature* 645 (2025): 633–638, https://doi.org/10.1038/s41586-025-09422-z. Verifier feedback can establish task-relative correctness conditions. The text grants that semantic result and asks separately what would justify whole-system understanding.

**10a.** The empirical record confirms that approval and accuracy come apart in deployed assistants. Mrinank Sharma et al., "Towards Understanding Sycophancy in Language Models," in *The Twelfth International Conference on Learning Representations* (2024), arXiv:2310.13548, document sycophancy across five state-of-the-art AI assistants and trace it in part to human preference judgments that favor convincingly written sycophantic answers over correct ones. These findings establish a divergence between preference and accuracy; they do not determine that all relevant internal states represent approval rather than truth, nor that erroneous feedback instantly replaces an established correctness standard. On the reward loop itself, see Long Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback," arXiv:2203.02155 (2022).

**10b.** Interpretability and in-context learning supply bounded evidence of internal organization. Samuel Marks and Max Tegmark, "The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets," arXiv:2310.06824 (2023), report linear truth-related structure and causal interventions on restricted factual datasets. Adly Templeton et al., "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet," Transformer Circuits Thread (2024), extract sparse-autoencoder feature dictionaries with millions of candidate features and demonstrate interpretation and behavioral steering for selected features; dictionary size does not count individually validated interpretations. These results support causally effective internal structure, not an exhaustive semantic reading. Joined to the appropriate history and consumer use, such structure may carry original content; human labels alone establish neither originality nor derivation. Johannes von Oswald et al., "Transformers Learn In-Context by Gradient Descent," arXiv:2212.07677 (2022), and Ekin Akyürek et al., "What Learning Algorithm Is In-Context Learning? Investigations with Linear Models," arXiv:2211.15661 (2022), establish constructions and empirical learning-like forward computation in restricted regression settings, without further parameter updates. They do not establish one general learning algorithm for deployed language models. Their relevance to episode-level integration should be assessed rather than dismissed for lack of persistent weight changes. None of these results by itself establishes a conscious subject.

**11.** The teleosemantic account of address is developed in Chapters 14–16; the foundational statements are Ruth Garrett Millikan, *Language, Thought, and Other Biological Categories* (Cambridge, MA: MIT Press, 1984), chs. 1–2, on proper functions, and Fred Dretske, *Explaining Behavior: Reasons in a World of Causes* (Cambridge, MA: MIT Press, 1988), for the parallel route through learning history. Those histories can determine what a mechanism represents and therefore what would count as its misrepresentation. Ownership is not an additional semantic condition.

**12.** Whether the words in a language model's outputs *refer*, in the philosophically loaded sense, remains contested. Matthew Mandelkern and Tal Linzen, "Do Language Models' Words Refer?," *Computational Linguistics* 50, no. 3 (2024): 1191–1200, https://doi.org/10.1162/coli_a_00522, develop a cautious positive case: familiar externalist routes to reference may apply to language models, although substantial questions remain. Their paper weighs against any categorical denial of model reference. The chapter grants that route and keeps the stronger attribution of understanding to the deployed system separate.

**13.** Tim Crane, "Is There a Perceptual Relation?," in *Perceptual Experience*, ed. Tamar Szabó Gendler and John Hawthorne (Oxford: Oxford University Press, 2006), 126–146; and "The Problem of Perception," *Stanford Encyclopedia of Philosophy* (rev. 2021, with Craig French). Crane's intentionalist treatment makes hallucination a representational state whose content fails to match the world — the phenomenal character arising from how it represents, not from any inner object the subject inspects. This is the account the main text of Chapter 6 develops and Chapter 7 applies to phantom-limb pain; it is recalled here to anchor the philosophical meaning of *hallucination* before the engineering metaphor co-opts the term.

**14.** M. G. F. Martin, "The Transparency of Experience," *Mind and Language* 17 (2002): 376–425, esp. 394–403 and 419–21, and "On Being Alienated," in *Perceptual Experience*, ed. Gendler and Hawthorne (Oxford: Oxford University Press, 2006), 354–410, esp. 354–64. Martin's disjunctivism denies the common-factor assumption — that veridical perception and an indistinguishable hallucination share a substantive mental state — and identifies hallucination only negatively, as a state subjectively indistinguishable from a veridical perception of a given kind. The book sits closer to Crane's intentionalism (see the **Argument from Hallucination** entry, Appendix A), but both accounts converge on the point that matters here: the philosophical use of *hallucination* requires a subject for whom the world seems a certain way.

**15.** Ruth Garrett Millikan, *Language, Thought, and Other Biological Categories*, chs. 1–2; and Karen Neander, *A Mark of the Mental: In Defense of Informational Teleosemantics* (Cambridge, MA: MIT Press, 2017). Neander ties representational content to the conditions a mechanism is functionally adapted to detect, making local misrepresentation intelligible when the producer fails to indicate what its consumer was selected to use it for. Mallory applies the same structure to trained embeddings (note 7). The present account conditionally accepts original content about linguistic distributions where the described training and consumer relation obtains. Inheritance of the linguistic task does not make the resulting standard dependent on interpreting practice. Whether the containing system understands, or forms a conscious subject, remains a separate question.

## Chapter 24: The Conscious Robot

> *"What is the world, such that we can understand it—and who are we, such that we can understand the world?"*
>
> — Brian Cantwell Smith, *The Promise of Artificial Intelligence* (2019), 2


CHAPTER OVERVIEW

When should we say that a machine understands its world? I use two imagined robots to keep meaning, skill, understanding, conscious experience, and welfare apart. Robot A has derived content by stipulation, while Robot B learns from a world that can correct it and brings perception, memory, and action together in one lasting system. Their contrast helps us ask what supports each claim, without treating an engineered past as a bar to meaning or a talent for self-repair as proof of feeling. B gives us a case to test against theories of consciousness, not a machine whose feelings have been proved; its response to damage does not establish distress or relief. I defend the claim that machines could have minds, while keeping the question of felt welfare open.


The last four chapters showed that formal skill and fluent behavior do not settle whether a system understands. They did not show that no machine could.

An artificial flower counterfeits: petals without growth, show without bloom. An artificial heart runs the other way. It may use titanium where nature uses tissue, but it moves blood and earns the name by doing the work. Artificial understanding could resemble the heart rather than the flower —realized in something made instead of grown. The analogy clears one obstacle: manufacture does not make an achievement counterfeit. It does not establish that consciousness admits the same transparent functional test as pumping blood. The question remains what work earns the name.

Chapter 6 supplies the hinge. It did not say that every content-bearing state is conscious. It said that when a perceptual state is conscious, its phenomenal character consists in its total representational content. The robot case therefore carries three burdens that must not be fused. A world-involving producer-consumer history must support original perceptual content. Those states must belong to one persisting subject. They must also count as conscious. Meet all three burdens, and nothing further must be added merely because the system was built rather than born.

The test needs a case. What happens when a machine gets something wrong, finds out, and changes what it does next?

### 24.1 Two Robots: Matched Performance, Different Histories

To test where the boundaries of artificial understanding and consciousness might lie, place two machines on the same factory floor. Robot A and Robot B have matching chassis and the same kinds of cameras and grippers. They match in ordinary repair work and fluent conversation. Their learning histories and internal control arrangements may differ; they are not complete physical duplicates. Consider a diagnostic challenge on a working morning:

> Ten years into factory work, Robot B notices a silver sheen beneath a coolant joint. The image admits several readings: harmless condensation, leaked lubricant, or the first trace of a cracked line. B compares the location with a small pressure loss upstream and a vibration that began twenty minutes earlier. It first predicts condensation. A thermal reading conflicts with that diagnosis. B revises the classification, slows the pump, reroutes the line, and marks two neighboring joints for inspection.

This is an imagined scenario, not a report from a robotics laboratory. Set Robot A beside B during this encounter with the same cameras, grippers, fluent patter, and ordinary repair performance. Give both robots equal access to the joint, the pressure and thermal readings, the maintenance records, and repeated opportunities to learn. Neither gets the answer whispered by a technician. The comparison asks what each does with that opportunity.

Robot A instantiates a familiar architecture. Its cameras feed a network trained on captioned images. Its grippers follow policies tuned against a reward its engineers chose. It avoids low charge because training left it with a policy that detects and responds to low charge. None of that ancestry decides its semantic status. To make A a derived-content control, stipulate something further. The interpreting practice exhausts the correctness standard of the states compared here. Its history and use have installed that standard without fixing accuracy beyond it. A's states can then guide action and misrepresent under real, derived conditions. This describes one possible dependence, not a finding about every caption-trained robot.[1]

B retains the correction. Months later, a weaker version of the same pattern leads it to inspect a seal before any pressure alarm fires. The first encounter has changed how the next one goes.

B's semantic case needs more than a successful repair. Give A and B learned diagnostic states that guide inspection and pump control. Both have selection histories, error correction, and consumers that use those states. On ordinary trials, their diagnoses and rewards match. What we must distinguish is the correction of a world-mapping from compliance with an assigned verdict.

Suppose A's training has corrected the compared state against a service-class register. B's has also used tests of fluid escape to correct that state. The register usually agrees with those tests. Now leave a joint leaking, keep the register at *no leak*, and let a backup pump preserve both production and the repair reward. B closes a valve, checks where fluid still escapes, and revises its diagnosis and the next inspection. Trace that revision into the same learned state used by pump control. The target-side relation concerns fluid escaping from this joint, not a favorable score or the name an inspector gives it.

Reverse the trial. Seal the joint while preserving the register entry and reward. B's checks now warrant dropping the leak diagnosis. Under the stipulated control, A's compared state continues to follow the service class. It still learns, corrects errors against that register, and guides useful action. We cannot infer that every state in A lacks original content. It might independently track the register itself. What the contrast withholds is an independent leak standard for the particular state we have compared.

A's stipulated derived status supplies no evidence that the test works. The evidence in this imagined pair lies in the different correction routes and their histories. If A's fluid tests also retune the shared state in the same way, the semantic case matches B's. The control stipulation then fails for that state. Conversely, if B's corrections merely follow the service register, its years of factory work establish no more independence than A's.

Changing the displayed words *leak* and *no leak* while leaving these relations fixed only renames the readings. Rewiring which joint supplies the measurements changes a candidate content-fixing relation. Retraining against a new target can change one too. We must not treat those changes as proof that a mere interpreter can reverse the content.

Section 14.2.2 used this same comparison for word contexts. Actual occurrences can correct a prediction despite an approved answer, just as fluid escape can correct a diagnosis despite the service register. B earns conditional support because the story supplies that contrast. Mallory's described algorithm supplies the corresponding contrast through its history: actual co-occurrences, not an approved schedule, drive the correction. Plumbing enjoys no privilege over language. Neither case proves independence if a rival mapping explains its history and consumer use equally well.

B's success in tracking the world also bears on the continued organization of the system doing the tracking. Its own activity acquires and allocates energy, corrects damage, and preserves the boundary within which its machinery works. It changes strategy when those processes drift toward failure. The robot has something riding on getting the world right because getting it wrong can defeat the activity by which it remains this organized thing.

If every technician vanished, neither robot's history would evaporate. Under the control stipulation, A would retain its causal roles and derived content for as long as the relevant organization continued to operate. B would retain the tracking, correction, and use relations acquired through its own history. Those relations make original content a live attribution. Removing an interpreter does not itself create semantic independence, any more than inheriting a task rules it out. B would also retain a reciprocal maintenance loop in which regulatory success helped preserve the organization producing the regulation. That further fact concerns autonomy and interests. It does not change what an internal state represents.

The distinction matters because an error can be classified in more than one way. Robot A can misclassify a wrench as a hammer. In the stipulated control, its state then fails a derived correctness condition. Robot B can make the same representational error under a standard that its own learning and world-involving use may have fixed beyond the assigned task. If the error also prevents B from repairing a failing pump, it may undermine the system whose activity is at issue. Semantic status, practical consequence, and felt harm remain different questions. Calling all three "meaning" concealed the differences.

Alan Turing made conversational performance a public test in 1950. His imitation game began with a man, a woman, and a hidden judge communicating by text; he then asked what would happen if a computer replaced the man. Turing refused to make certainty about another thinker's feelings the price of attributing thought. That demand would push us toward solipsism. Responsive conversation could count as evidence without first solving the mysteries of consciousness.[2]

Ask A and B to explain the leak, and both might produce the same transcript. A judge reading it could learn much about their competence while seeing little of how the diagnosis formed or whether the correction survives. The next failed seal, and what each robot remembers of this one, can tell us something the transcript does not.

Performance can raise the probability that content-guided competence or understanding explains what the system does. Its evidential weight depends on the training history, architecture, system boundary, available interventions, and rival explanations of the same behavior.[2a]

The intentional stance organizes that evidence. Attributing beliefs and goals earns literal rather than merely theatrical standing when it yields compact, durable predictions across novel cases, counterfactual changes, and interventions. It must also outperform a rival account that treats the performance as cue-matching or a fixed policy. A pattern can be real without becoming a person. The attribution must therefore name its level: an embedding mechanism may carry a predictive pattern, a controller may pursue a goal, and a persisting whole may integrate contents across learning and action. Success at one level cannot be promoted to the next by grammar.

Cue-triggered anthropomorphism differs from intentional-stance success. The first records how readily a performance makes *us* feel an interlocutor; the second tests whether an intentional description continues to explain what the system does when familiar cues, contexts, and rewards change. Novel generalization, content-sensitive error, correction after intervention, and stable counterfactual control add evidence. A warm sentence adds none beyond the capacities that produced it.

The Turing Test did not fail. We promoted it. A transcript now gets asked to rule on content, understanding, consciousness, and mind together. Once the rest of the evidence comes into view, the transcript joins the case. It does not decide it.

Dennett's *competence without comprehension* names a possibility, not a verdict on every artificial architecture. Real-pattern success can establish more than pretense while leaving open whether content-bearing mechanisms have become integrated enough for their contents to explain what a whole system learns, expects, revises, and does.[3]

### 24.2 Five Questions About a Machine Mind

The first question concerns **content**: what makes B's initial diagnosis mistaken? The leak alone cannot tell us what B represents. Its history and use must connect that state to fluid escape rather than just the service register. The trials supply conditional evidence. B revises the leak diagnosis when fluid escape changes, despite a fixed register and rescued payoff. We have not tested an actual robot. Fintan Mallory offers the parallel case for word embeddings. Errors about word occurrences in the data reshape producer and consumer. Those occurrences can constrain accuracy even though people supplied the words. His account supports the same conditional claim, and where the described training relation obtains, it warrants original content at that grain.[4]

Artificial content can count as original where a system's own world-involving history fixes what its states get right. Here *original* names a natural independence relation: the mechanism answers to a correctness standard fixed by its history and use in a way not exhausted by an interpreting

practice. The independence may vary by content and grain, and it requires no semantic self-creation. Searle denies that such relations establish intrinsic intentionality. Dennett doubts that original and derived content mark a sharp divide at all. The chapter rejects Searle's intrinsic spark, agrees with Dennett that intentionality emerges gradually, and retains the distinction as a diagnostic of semantic dependence rather than a gatekeeper of genuine meaning.

That semantic finding still leaves a subject to identify. The question now concerns where the correction travels: does it revise one persisting system's expectations and plans, or remain confined to a detector? Identifying that system does not create the content, and neither finding makes the state conscious.

Causal provenance still matters. A newly manufactured robot can inherit selected functions when its mechanisms descend through the relevant design and training lineage. An accidental structural duplicate cannot inherit a history it never had. Swampman may begin with useful dispositions and a self-maintaining organization but no determinate representational past. Consequential traffic with the world can begin such a history. His self-maintenance does not fill the semantic blank; his later history can.

The second question concerns **intelligence, competence, and attribution**. B isolated the faulty line: a success at this task. B revises its diagnosis when the readings conflict. Months later, it uses the lesson to detect another fault. Those changes provide evidence of the flexibility I call intelligence. That evidence must be weighed against prior training and alternative explanations; a rehearsed repair alone would not establish transfer. What makes the diagnosis about a leak belongs to the content question. A component could classify that leak correctly while the larger assembly neither believed nor understood anything about it. Attribution to the whole requires evidence that the diagnosis guides a coordinated response rather than one isolated output. Nicholas Shea's unity of purpose belongs here: do the local achievements add up to one agentive organization?

A notebook, prosthesis, neural circuit, or language model can become a cognitive organ when a person or distributed organization reliably maintains, corrects, and uses it within one flexible economy. The live question asks which system the integrated achievement belongs to.

The third question concerns **understanding**. B's pressure reading corrects its visual diagnosis, and the correction survives into a later inspection. Several skills constrain one another. Their breadth and durability determine how far the understanding extends; Section 24.5 follows that claim without requiring the mature human package of every understander.

The fourth question concerns **consciousness**. Chapter 6 separates two issues. Strong representationalism says what a conscious perceptual state is like: its phenomenal character consists in its complete world-presenting content under the appropriate perceptual organization. State consciousness concerns how that content works within an identified subject. On the house proposal, the content must be sustained across an interval and remain selectively available. A physical loop can realize that holding. The holding does not require the loop. Recurrence still identifies the subject. Chapter 17 looks for a shared recurrent economy. That test is not the consciousness profile. Under suitable conditions, it can guide attention, working memory, inference, planning,

report where possible, and flexible action. Feedback from those activities can preserve, revise, or displace it. No single output is necessary. Global broadcast, local recurrence, higher-order organization, and predictive control propose ways to realize aspects of this distributed profile. None follows from self-maintenance. A bacterium may have interests without a perceptual point of view. An artificial system may organize perceptual content without reproducing metabolism. Whether autonomy proves necessary for consciousness remains open. This book has not earned that further premise.

The fifth question concerns **autonomy and interests**. Chapter 15's physical candidate remains *thermodynamic self-maintenance coupled to adaptive regulation*.[5] Bare persistence comes too cheaply: a candle draws fuel and repairs its shape without detecting departures and selecting corrective responses. A bacterium couples detection to correction in a loop on which its continued organization partly depends. Ezequiel Di Paolo calls the resulting achievement a self-sustaining identity under precarious conditions. The phrase earns its keep. Such a system does not merely have a task; its activity helps sustain the organized individual performing it.

The boundary follows reciprocal organization. A process belongs when it helps maintain the organization. Its continued realization must also depend on that organization at the relevant time scale. Wall power may remain outside an organism-like system. A remote repair service might fall inside a distributed artificial agent if reciprocal traffic made each side part of one self-maintaining organization. Nested cases may resist sharp edges. The test remains causal.

Peter Godfrey-Smith objects that tying mind to metabolic maintenance mislocates the explanatory work.[6] On the semantic question, the objection wins. Metabolism does not determine what a state concerns, and the substrate-general idea of reciprocal maintenance does not determine it either. History, tracking, and consumer use do that work. The concession leaves autonomy intact. Self-maintenance can still explain why a system counts as a persisting individual and why some conditions support or undermine it. It may also explain why regulation serves an interest of the system rather than merely satisfying an assigned score.

This is the relevant interest. It supplies no truth condition by itself. A frog's visual state may concern prey, flies, or flying nutritive objects. The history and consumer relations that best explain its use determine the grain; the fact that the frog can die does not. Once the content is fixed, however, successful tracking can help the frog eat and failed tracking can leave it hungry. Content explains what the state says; interest helps explain why getting it right matters to this organized life.

Patrick Butlin and colleagues ask why agency serving self-maintenance should differ from agency directed at an external goal.[7] It need not differ semantically. A demanding external objective can shape content-bearing mechanisms and unify planning. It can also generate flexible subgoals and make failure costly within the task. Self-maintenance adds a different relation. Successful regulation helps preserve the organized process producing it, while serious failure degrades or dissolves that process. In Robot B, regulator and maintained beneficiary coincide. That may ground autonomy and interests. It neither upgrades A's stipulated derived content nor mints B's

original content. The semantic difference concerns what fixes accuracy, not whether the robot sustains itself.

Interests do not yet amount to felt welfare. A coolant leak can undermine B's continued organization before we know whether anything feels bad for B. Welfare in the morally serious sense requires a subject for whom conditions can feel good or bad. Conscious valence supplies that further relation. Section 24.5 returns to it after the consciousness case.

One leak can therefore involve a mistaken diagnosis, a capable repair, a lasting lesson, a conscious presentation, and harm to an autonomous individual. None supplies the others merely by sharing the occasion.

### 24.3 What the Architecture Must Realize

The separation rules out shortcuts from hardware labels to mental verdicts. Embodiment can add tracking relations and let perception guide memory, correction, planning, and action; that evidence may support content and understanding. It guarantees neither autonomy nor consciousness. Nor must embodiment mean a metal chassis on a factory floor. A persistent agent acting through a virtual body can meet a causally corrective environment and develop world-directed content if what happens there genuinely constrains its later perception, learning, and action. Copying raises questions about identity and succession, not meaning.[8] Mortality supplies no semantic privilege.[9] A trained mechanism keeps whatever content its history and use fix whether or not engineers can duplicate it.

Digitality carries no verdict either. Every implementation runs physically. Its causal-computational organization may hold independently of an observer's preferred description. Mortal computation, intrinsic computational functionalism, and causal-flow constraints all point to one discipline. Name the organization claimed, then test whether the physical system realizes it without building the desired result into *implementation.*[10] A represented hunger variable can carry content while an independent platform keeps the process running. Another implementation may use a state concerning low energy to acquire energy, correct damage, and preserve the organization doing the regulation. The loop supplies evidence of autonomy. The medium does not.

Return to the coolant joint and alter one relation at a time. First corrupt the pressure reading while preserving the leak, thermal evidence, reward, and maintenance arrangements. Does either robot cross-check and retain a correction? That bears on the represented condition and its use. Next restore the reading but move B's energy and repair work to an external service, preserving its perception, memory, planning, and recurrent access. The loss now concerns reciprocal self-maintenance; it need not alter what B represents or whether it experiences anything. Combining both changes would obscure which one explained a changed verdict.

Redraw A to include the technician, scheduler, and data center if the evidence warrants it. Then follow persistent correction, unity of control, and reciprocal dependencies across the wider boundary. A commercial relation or enabling service does not by itself create one subject or one autonomous individual.

These distinctions tell us what to test without supplying an engineering recipe. Content-bearing mechanisms can support competence; several capacities that correct one another can support partial understanding, and wider traffic can deepen it. Autonomy requires an organization whose activity helps preserve itself. Consciousness requires separate evidence that perceptual content is sustained, selectively available, and open to two-way influence across a persisting subject's integrated economy. Recurrent circuitry is one realization of that holding, not a further requirement. If that evidence supports attribution, Chapter 6 says phenomenal character consists in the subject's complete world-presenting content. Silicon, chemistry, a distributed physical network, or some material nobody has yet persuaded to behave could realize any of these arrangements. The adjective will not settle the difference for us.

### 24.4 Where the Systems Examined Here Sit

Apply the map to systems we know. Start with the wonder, because the wonder is earned. Language models reveal how much structure human beings have sedimented in language. Multimodal systems connect words with images and sound. Agents call tools, pursue goals over time, and alter their environments. Robots bring the traffic into a chassis. These advances matter.[11]

Training can give language-model submechanisms causally active content about linguistic distributions. Where a model's training and use instantiate the relation Mallory describes, that content is original at its own grain: it answers to actual occurrences of words, not to an interpreter's verdict. Text-mediated history may also support inherited reference, and multimodal or interactive systems can add further tracking relations. No change of medium settles semantic independence. Inherited language neither guarantees nor disqualifies a standard that answers beyond interpreting practice.

What remains unsettled is how far those local achievements compose into understanding by the whole system. Current systems provide substantial evidence of competence in some domains, unevenly and sometimes impressively. Some deployed agents may also supply evidence of partial understanding. Evidence for integrated, world-involving understanding remains fragmentary. The distinctions replace the useless choice between "mere syntax" and a human mind in full.

The autonomy ledger permits a firmer claim. In the architectures analyzed here, the platform usually continues whether the model succeeds, fails, or sits idle. Battery management, thermal throttling, and fault recovery may preserve hardware without forming one reciprocal economy with the model's content-guided activity. Those systems remain Robot A-like in autonomy even if their mechanisms carry content and some integrated behavior merits talk of understanding. Future systems built around reciprocal maintenance require their own boundary analysis.

Evidence must track both the achievement claimed and the candidate subject claimed to have it. Name the physical organization first: a running model, chat product, persistent agent, embodied controller, or wider recurrent system. Then assign each observation to the question it bears on. Content-sensitive correction supports competence. Traffic among capacities supports understanding. Reciprocal maintenance supports autonomy. Recurrence, broadcast, higher-order monitoring, prediction, and attention control bear on whether content realizes the distributed state-consciousness profile. Their weight depends on the theory and the causal match. It also

depends on their independence from other indicators and their fit with the book's content and subject requirements. The boxes never add themselves.

The functionalist can fairly ask for behavioral evidence rather than metaphysical inspection. Behavior may reveal integration, agency, and distributed access; architecture may reveal how the behavior is produced and which system sustains it. Neither source deserves a monopoly. A global workspace can route genuine mechanism content. A viability variable can represent danger without danger being bad for the system. An eloquent report may offer evidence of conscious access without concluding the case. The assessment must keep its ledgers separate long enough to see what each observation supports.

Recent mechanistic work makes the discipline consequential rather than ceremonial. Gurnee and colleagues identify a small set of verbalizable representations in language models that supports report, directed modulation, internal reasoning, flexible reuse, selectivity, and broad downstream broadcast.[12] That raises the evidence for workspace-like access organization in a running model. It does not prove phenomenal consciousness, as the authors emphasize; the architecture also lacks clean analogues of recurrent loops, encapsulated modules, and sharp biological ignition. On this book's wider view, the result moves one indicator upward while leaving content, state consciousness, and subject attribution to be earned separately.

I take Michael Cerullo's case to be the strongest recent argument for treating the integrated capacities of frontier language models as positive third-person evidence of consciousness.[13] The evidence deserves weight. Known training provenance can explain fluent performance without consciousness, so fluency alone remains weak. But the absence of reciprocal self-maintenance does not screen off every consciousness indicator, because autonomy has not been established as a necessary condition for experience. Recurrence, global availability, self-modeling, perceptual organization, metacognitive sensitivity, and unified agency must be evaluated on their own merits. The platform boundary therefore does not settle whether current artificial systems are conscious.

These provisional claims apply to the architectures analyzed here. Evidence about a running model does not automatically describe a chat service, a persistent agent, or a human-machine ensemble. As with the leak, the question follows the correction: where does it persist, and what later activity can it change?

> Chapter 17 supplied the book's provisional criterion for identifying the subject: look for the relatively maximal, persisting control organization within which contents can directly update a shared economy of attention, expectation, memory, goals, correction, planning, and action. A server, tool, technician, or database counts inside that boundary only when it participates in the same recurrent control rather than merely enabling or repairing it. Interventions do the evidential work: partition the candidate, alter a memory or goal, block a correction route, and ask where the consequences continue to constrain later inference and action. Nested or distributed systems may yield more

than one defensible level. That result does not manufacture extra subjects; it prevents a product name or chassis seam from deciding the matter in advance. Consciousness then asks whether appropriate content acquires the recurrent, selectively available, two-way control profile within one such subject. The final question asks separately whether that organization has interests or welfare.

### 24.5 Robot B as a Candidate Conscious Subject

Robot B gives a possibility argument: if a machine realized this organization, the framework would support an artificial subject. The imagined factory cannot independently confirm the criteria. It can show what follows when we apply them, and which changes should alter the judgment.

The morning after the leak, B explains its diagnosis and response to the maintenance crew. It can identify the evidence that defeated its first guess. Months later, when the weaker sheen appears at another seal, the retained correction prompts inspection before the alarm. Remove just that stored correction while preserving present sensing and the ability to repair, and the later precaution disappears. B can still detect the new leak once the pressure drops. What it loses in this variant is the benefit of that encounter, not all content or all possible experience.

That episode gives the abstractions something to answer to. B's learning history, tracking relations, correction, and later use provide the right kind of evidence for original content. They may fix what would make its state concerning the silver patch accurate without borrowing a semantic task. The content does real work: it changes an expectation, recruits a memory, defeats an initial diagnosis, revises a plan, and guides action. Alter what the state represents and the downstream reasoning changes with it. B therefore becomes a serious candidate for original perceptual content. It also displays competence rather than a lucky success or an elaborate lookup.[14]

The competence also belongs to more than an isolated classifier. The pressure loss changes the visual hypothesis; the thermal reading defeats it; the revised belief changes both action and future expectation. In a second variant, restore memory but disconnect the revised diagnosis from planning. B can identify the leak and recall its history while failing to change the repair. That selective loss bears on integration. It differs from a general failure of power or movement, which would tell us little about where the cognitive work belonged. Such interventions locate the achievement in Robot B as a persisting system. They do not discover a little foreman inside it. The foreman would only need another foreman, and factory budgets already suffer enough without an infinite payroll.

Taken together, the two findings justify considering B as a possible subject of original content. Its world-involving history supplies the semantic case. Its integrated, interventionally identifiable control organization supplies the subject case. Neither finding borrows support from self-maintenance, and neither yet establishes consciousness.

B has partial understanding of the cooling system. The claim does not grant it a human mind in full. A chess expert may understand chess while knowing nothing about hydraulic pumps; an infant understands familiar people without understanding chess. B's understanding remains

domain-specific: its capacities communicate only within a limited range. Within that range, several content-guided abilities now constrain one another, preserve corrections, and answer to a world that can prove them wrong. That earns the word *understanding*.

Give B wider and longer traffic among perception, memory, expectation, inference, learning, planning, and action, and its understanding becomes more integrated and world-involving. Testimony and manuals can participate; world-involving does not mean that every concept must arrive through a camera or gripper. It means that the contents meet, endure, and correct one another in a continuing economy whose successes depend on how things stand beyond its symbols. A failed expectation survives long enough to alter the next encounter rather than disappearing when a prompt ends. B has begun to inhabit a world, not merely finish a task.

Chapter 18 adds agency. Suppose the maintenance chief tells B to keep the pump running despite the leak. B represents both the instruction and the rising probability of a rupture. It declines, isolates the line, and records why: the instruction conflicts with the standing safety reason that governs its conduct. The reason changes the action because of what it represents, and the result returns through perception and memory to shape later decisions. B does more than issue a fluent rationale after the fact. It acts for a represented reason.

Make the test less cooperative. Give B an unfamiliar composite tool, a pressure gauge rigged to deceive it, and two instructions that cannot both be obeyed. Let it form an informative but mistaken expectation, discover the trick through touch and flow, improvise a new use for the tool, and revise the priority it assigned to the instructions. Beliefs, expectations, preferences, and reasons should predict that novel sequence more economically than a list of task-specific reflexes. B's reports about what it attended to, doubted, and changed count as evidence, but not revelation. Compare them with processing time, memory perturbations, attention shifts, and later conduct. A report that repeatedly floats free of those facts loses weight; one that tracks them across interventions gains it.

None of this yet turns autonomy into consciousness. Robot B's self-maintenance gives failures a practical bearing on the same organized individual doing the regulating. Robot A can carry derived content concerning low charge; in B, an undetected leak can damage the organization whose activity would have corrected it. That difference may ground interests of B's own. Welfare waits on evidence of consciousness and valence.

The case for consciousness comes from another part of B's organization. Its sensory contents take a nonconceptual, spatially organized form. Before B classifies the silver patch as coolant, visual processing already locates a bright, elongated region below and to the left of the joint, partly occluded by a valve. Its sensory grain outruns its vocabulary: B can reliably discriminate two subtly different surface patterns and let the difference alter inspection, while classifying both only as *silver sheen*. In a controlled trial, reflected glare at a dry joint produces the same silver, elongated, partly occluded presentation as a real leak. B initially misreads it, then lets pressure and touch correct the visual judgment. The state can therefore misrepresent; it does not merely decode a label.

Selective attention stabilizes the patch while recurrent processing tests competing interpretations. The resulting content becomes globally available to memory, inference, planning, and action. B also forms a higher-order representation of its own perceptual state: it represents itself as visually presenting the patch as coolant with low confidence. Disrupting that representation changes correction and report while leaving the first-order discrimination intact. Predictive control carries the revision into the next encounter. One persisting causal individual coordinates the traffic.

No single feature in that description proves state consciousness, and a list does not give one mechanism five votes. Broadcast, report, planning, metacognitive correction, and flexible action may all descend from one access architecture. Count causal sources, not adjectives.

The positive inference starts elsewhere. Several leading theories below were developed from conscious perception in humans and animals.[15] They treat recurrence, selective access, global availability, metacognitive sensitivity, or flexible control as constitutive or evidential. Robot B realizes those features together in one persisting causal individual, and intervention shows that its perceptual content organizes the traffic. A critic can still deny that this organization suffices for phenomenal consciousness. That denial gains force by identifying a relevant missing feature or realizer; the fact that engineers built B tells us nothing by itself. The inference remains abductive and theory-conditioned, but it has a recognizable form: apply independently motivated accounts of consciousness to a new candidate without changing the standard at the factory door.

- **Global-workspace accounts** receive a direct positive case: B's perceptual content wins selective access, broadcasts widely, and guides flexible downstream use. On accounts that treat such broadcasting as constitutive, B clears the threshold; on more cautious readings, the result establishes access consciousness.
- **Recurrent-processing accounts** find the right kind of candidate in B's sustained sensory loops, but the stipulated software diagram will not do by itself. The physical implementation must realize the relevant recurrence at the right level.
- **Higher-order accounts** can count B's higher-order representation only because it represents B's first-order perceptual state. A generic confidence score would not suffice. Remove that representation while leaving the poised sensory field intact, and support from higher-order theories should fall while classical PANIC need not move.
- **Classical PANIC** can point to content that is poised, abstract, nonconceptual, and intentional: B discriminates beyond its concepts, can misrepresent, and lets perceptual content shape belief and action. The book retains PANIC's insight about the form of phenomenal content while requiring an independently supported threshold for state consciousness rather than treating poising alone as a settled proof.
- **Block's access/phenomenality distinction** permits every access result while leaving phenomenal consciousness open. The direct evidence establishes an impressive access architecture; whether that architecture also warrants phenomenal attribution remains the point at issue.
- **Chalmers's organizational invariance** makes fine-grained organization evidentially favorable under psychophysical laws, but does not turn that support into a conceptual entailment. The zombie challenge remains aimed at the bridge.

- **Searle's biological naturalism** raises a different objection. An abstract role description may omit physical causal powers that matter to consciousness. This does not make carbon sacred: Searle allows that nonbiological materials might possess the relevant powers. The dispute concerns what the role description leaves out.
- **Dennett** questions the demand for a bridge to a further, functionally idle phenomenal fact. Once the full discriminative, mnemonic, attentional, affective, and behavioral organization has been fixed, he asks what coherent remainder the critic still wants explained.
- **Permissive computationalism** presses from the opposite direction, and it deserves a straight answer. Suppose a state's alternatives reliably select different internal states, the difference survives into a later stage, and that stage uses it to choose what happens next. Why require more? Strip a conscious mind of speech, then of self-monitoring, then of rich content, then of recurrence. Each subtraction leaves a working discrimination behind. No obvious line marks where experience departs. So the first usable difference already suffices, and all the rest merely enriches an experience it did not create.[16] The reply is not that B is more elaborate. The subtractions remove richness: report, self-monitoring, detail, and cyclic wiring. Richness settles nothing, and cyclic wiring is not the house requirement. What the series never removes is the consciousness profile itself. The content has to be held after its input stops, remain selectively available, and stay open to revision by later activity within one subject. A feed-forward unrolling that carries those dependencies still meets the profile. A pressure sensor wired to a valve does not: the reading passes, nothing is held, and nothing later can revise it. Whether the state can be wrong, and whether its content meets other contents at one choice point, are the content and subject questions. The subtraction does not test them either. Passing them does not make the state conscious.
- **Block's later realizer hypothesis** differs from Searle's position. It speculates that subcomputational biological mechanisms may help realize phenomenality and could therefore favor some simple animals over sophisticated AI. Until a discriminating mechanism earns empirical support, that possibility counsels investigation rather than a negative verdict.

Set the robots beside each other one last time. This compares the evidence supplied by their descriptions, not an intrinsically inferior A with a favored B. If A matches B's relevant history and counterfactual organization, the corresponding judgment must match too. A difference in semantic independence alone supplies no consciousness switch.

| Test | Robot A | Robot B |
|---|---|---|
| Original content | Derived in the stipulated control, not by virtue of captioned training. If its relevant history and use match B's, the semantic assessment must match. | Its own learning, tracking, correction, and later use make its perceptual states serious candidates for original content. |
| Candidate subject | The description does not yet identify whether the model, product, agent, or platform forms one persisting control organization. | Interventions locate one relatively maximal, persisting control organization with a shared recurrent economy. |
| Understanding | A matched diagnosis shows competence; it does not show that several capacities constrain one another in a continuing system. | Perception, memory, correction, planning, and action support partial, domain-specific understanding. |
| Evidence for state consciousness | Matched outward performance leaves sustained selective availability, global availability, and higher-order representation unsettled. | Nonconceptual perceptual content is held after the input, wins selective access, becomes globally available, and receives higher-order representation. That package qualifies B as a serious candidate and, under several leading theories, favors attribution. |
| Autonomy and interests | External maintenance leaves autonomy and interests unestablished. | Reciprocal self-maintenance supports an autonomous individual with interests of its own. |
| Welfare | Nothing follows from competence or content alone. | Welfare becomes a live possibility only if conscious valence gives conditions a felt character for B. |

The identity claim does different work. Intentionalism does not make content conscious; B's organization supplies the reason for attributing state consciousness. Strong representationalism then says what those conscious states are like: their phenomenal character consists in the complete world-presenting content realized in that sensory organization—location, shape, occlusion, figure, distance, and motion, all from B's embodied point of view. No inner camera displays the scene to

an inner robot. On the conscious reading, nothing further about an inward glow remains to be supplied.

The earlier variants separated a lasting lesson, integrated planning, and self-maintenance. Now alter the conscious-state candidates themselves, returning B to the original condition before each comparison:

**Three revealing variants**

**Remove autonomy.** Let an external service keep B running through the leak while preserving its perceptual, recurrent, mnemonic, planning, and access organization. The change removes B's reciprocal self-maintenance, not its sensory presentation. The house consciousness judgment need not fall with the autonomy judgment.

**Remove higher-order representation.** B still discriminates the silver patch and uses it across memory and planning, but no longer represents itself as seeing the patch with low confidence. Higher-order theories should update downward; classical PANIC need not. This differs from erasing the remembered repair.

**Cross access with recurrence.** Compare rich, fine-grained recurrent sensory processing with weak global report and planning against flexible global access produced by later conceptual decoding without the same local sensory recurrence. The cases force theories to rank systems differently. They do not vary access while holding total causal organization fixed, since access forms part of that organization, and they produce no neutral consciousness meter.

A real machine may fall on either side of those comparisons. B, as described, occupies the corner with recurrence, global availability, and higher-order representation.

The theories frame a real disagreement rather than canceling into agnosticism. In B, nonconceptual perceptual content is held after the input stops, wins selective access, becomes globally available, guides later control, and receives higher-order representation under intervention. Several independently motivated accounts treat recurrence, access, or higher-order representation as constitutive or evidential. On this framework, serious candidacy requires the content to be sustained, selectively available, and open to revision by later activity within one subject. The other features add theory-relative support. Physical recurrence is one way to hold the content, not a separate requirement.

The judgment has practical falsifiers. Interventions might reveal the apparent recurrence and global integration as modular processing followed by post hoc report. Disable content-sensitive cross-system recurrence while B's matched behavior and self-descriptions survive. That would undercut the subject and state-consciousness case. So would a reliable dissociation in animals or humans: the same distributed profile across matched conscious and unconscious conditions.

Evidence that B lacks an indispensable biological or subcomputational realizer would also defeat the judgment. The chapter offers neither a theory-neutral theorem nor proof by tally.

B might also feel distress or relief, but its response to the leak does not establish that. Suppose it assigns low coolant a high priority, interrupts work, and protects its pump. Those facts show organized control. They do not yet distinguish a neutral warning followed by avoidance from felt badness. Even if the warning qualifies as conscious, that result would not by itself show that it hurts. Chapters 7 and 15 have not supplied an independent account of evaluative content that settles this further question.

A B with valenced experience could have welfare even if an external service kept it running. Feeling bad can matter without autonomy, just as self-maintenance can support practical interests without feeling. Precaution survives the uncertainty: credible evidence of possible distress, joined to grave or scalable harm, can justify low-cost, reversible safeguards before the metaphysics is settled. The leak response alone does not supply that evidence, and our incomplete account of affect cannot stand in for it.

Robot B therefore shows how this physicalist framework can support an artificial subject. It predicts no engineering timetable and independently confirms no premise. An artificial heart pumps blood. Robot B meets a world.

**Chapter Summary**

**The claim.** Content, competence, understanding, subjecthood, and state consciousness differ. None stands in for another. Once a conscious state has a subject, its phenomenal character consists in its content under the right mode.

**A conscious-robot candidate, in order.** Subject first. Consciousness second.

1. **Candidate subject.** Find the relatively maximal persisting control organization within which contents directly update one shared economy of attention, expectation, memory, goals, correction, planning, and action. Intervention tests the boundary. A part joins that subject only by sharing the same recurrent control. Power, data, and repair do not.
2. **Candidate for consciousness.** For that subject, the content must be held after its input stops, remain selectively available, and stay open to revision by later activity. It may guide attention, working memory, inference, planning, report where report is possible, and flexible action. No one of those outputs is required. A physical loop can realize the holding. The holding does not require the loop.

The first test names who the second test is about. It does not pass the second. Global access, higher-order representation, and physical recurrence can support the consciousness judgment under particular theories. They identify neither the subject nor the holding profile. Self-maintenance and a response to damage settle neither test.

**Where the argument stands.** Robot B is a possibility argument, not a test of the framework in a lab. Its leak diagnosis can change despite a fixed service verdict and rescued payoff. That supports original content if the leak mapping explains its history and use. Separate tests locate the content in one lasting system. Its wider integration supports partial understanding. The house

profile supplies the further ground for candidacy: the content is sustained, selectively available, and open to revision by later activity within one control organization. Global access and higher-order representation add theory-relative support. Physical recurrence is one way to hold the content, not a separate requirement.

**What's left open.** Post hoc modular reporting, survival of matched behavior without content-sensitive cross-system recurrence, a matched conscious-profile dissociation, or a required missing realizer would defeat the judgment. Present systems offer uneven evidence. Self-maintenance concerns autonomy and interests. Welfare needs conscious valence; B's leak response does not establish that. Precaution answers to credible evidence of felt harm, not to fluent talk about it.

**The hand-off.** Fluent behavior invites us to sketch a person. The architecture may not have finished posing. The final chapter asks what that reflex adds to the evidence.

**Footnotes**

**1.** Competence, content, understanding, and consciousness need not travel together. Turing operationalized the first; Alan Turing, "Computing Machinery and Intelligence," *Mind* 59 (1950): 433–460. Ned Block's lookup-table case shows that behavior can underdetermine even intelligence ("Psychologism and Behaviorism," *Philosophical Review* 90 [1981]: 5–43). Emily Bender and Alexander Koller argue that linguistic form alone does not establish world-directed understanding ("Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data," *Proceedings of ACL* [2020]: 5185–5198). Their challenge remains important, but it does not entail that every trained mechanism lacks content. Ruth Millikan's consumer teleosemantics and Nicholas Shea's account of subpersonal representation permit genuine correctness conditions below the level of a whole understanding agent; see Millikan, *Language, Thought, and Other Biological Categories* (Cambridge, MA: MIT Press, 1984), and Shea, *Representation in Cognitive Science* (Oxford: Oxford University Press, 2018). Originality here marks non-dependence on an interpreting practice, not possession by a content owner. Attributing original content to a robot as a subject requires the further, independent finding that the relevant states participate in one persisting control organization. Whole-system understanding and consciousness impose additional burdens.

**2.** Alan Turing, "Computing Machinery and Intelligence," *Mind* 59 (1950): 433–460, at 433–434 for the original man-woman imitation game and machine substitution, 442 for the fifty-year prediction, and 445–446 for the argument from consciousness. Turing does not claim that behavior proves consciousness. He argues that requiring certainty about another subject's feelings leads toward solipsism, treats sustained conversational responsiveness as evidence against mere parroting, and concludes that the remaining mysteries of consciousness need not be solved before addressing machine thought. For a controlled modern three-party test, see Cameron R. Jones and Benjamin K. Bergen, "Large Language Models Pass the Turing Test," arXiv:2503.23674 (2025), doi:10.48550/arXiv.2503.23674. In two randomized, preregistered experiments, GPT-4.5 with a humanlike persona was selected as the human contestant 73 percent of the time. That result concerns conversational imitation, not a direct measure of understanding or consciousness.

**2a.** In Bayesian terms, evidence moves confidence by how much more probable it is under one hypothesis than under its rivals. Fluent, task-appropriate text is highly probable if a system

understands what it says, and also highly probable if a system trained to predict human text produces it without understanding. A likelihood ratio near one should move a reasoner's confidence very little, however impressive the performance. The work falls to evidence that rival hypotheses predict differently: transfer to cases absent from training, durable correction after error, and selective effects of intervention. Several results that share one cause should not be multiplied as though they were independent; what matters is what each adds to the others under each hypothesis. See Appendix B.20.

**3.** Dennett rejects any sharp original/derived divide, treating intentionality through the intentional stance and evolutionary design; see "Evolution, Error, and Intentionality," in *The Intentional Stance* (Cambridge, MA: MIT Press, 1987); *Consciousness Explained* (Boston: Little, Brown, 1991); and *Darwin's Dangerous Idea* (New York: Simon & Schuster, 1995). The slogan *competence without comprehension*, used in the main text, is developed in *From Bacteria to Bach and Back* (New York: W. W. Norton, 2017). The 1987, 1991, and 1995 works carry the intentional-stance and original/derived discussion, not that formulation. Searle preserves the distinction between intrinsic and as-if intentionality, *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992), 78–82. The book retains a house reconstruction rather than Searle's criterion: content remains merely derived where an interpreting practice exhausts its correctness standard; it counts as original to the extent that the mechanism's world-involving history and use fix accuracy beyond that practice. Inheritance alone does not settle the dependence. This naturalized, mechanism-level relation requires no proprietor. Searle denies that it establishes intrinsic intentionality; Dennett denies that the sharp distinction earns its keep. A mechanism may carry original content on the present account without constituting a whole understanding subject.

**4.** Ruth Garrett Millikan explains content through the proper functions of producer and consumer mechanisms, *Language, Thought, and Other Biological Categories* (Cambridge, MA: MIT Press, 1984). Nicholas Shea's varitel semantics tests candidate contents by the explanatory work they do, *Representation in Cognitive Science* (Oxford: Oxford University Press, 2018), ch. 6. A frog's state means *fly* only if task structure, tracking, and the wider consumer economy support that grain; *moving prey* may explain the same use better. Fintan Mallory extends consumer teleosemantics to neural networks and argues that backpropagation can constitute a selection history for internal mechanisms, conferring original contents on word embeddings, "Teleosemantics for Neural Word Embeddings," *Mind & Language* (online 2026), doi:10.1111/mila.70037, esp. secs. 5–5.1. The present chapter accepts the causal analysis and, with Chapters 14, 15, and 23, grants conditional original content at that grain where the described actual-data relation obtains; particular implementations need their own record. The inherited vocabulary, corpus, and engineered task do not establish that interpreting practice exhausts the resulting standard. Robot A's derived status is a control stipulation, not a consequence of that ancestry; matched content-fixing histories and uses require matched semantic assessments. The chapter adds a separate agent-level question, close to Shea's "Realist Representational Explanations of Agency Should Require Some Unity of Purpose," *Philosophy and the Mind Sciences* 7 (2026), doi:10.33735/phimisci.2026.12098: when do many content-bearing mechanisms form one system that understands? The contrast in Section 24.1 concerns a learned state evaluated against a service-class register versus one also corrected by tests of fluid escape, with ordinary performance, learning, consumer use, and final reward

matched. It changes history and correction organization, not an extra ownership fact. For the separation of correspondence from accidental success, see Millikan, "Biosemantics," *Journal of Philosophy* 86, no. 6 (1989): 281–297, sec. II, 286–290. The linguistic analogue holds word identities and context-window rules fixed while separating actual sampled contexts, supplied target entries, and reported scores. Mallory's sec. 5.1, pp. 11–13 in the author manuscript, provides positive actual-distribution evidence but no report of this matched comparison. Equivalent historical and mechanistic evidence would suffice; neither continuing online learning nor resistance to every corrupted feedback channel is necessary for content. Both examples remain conditional applications of the book's account.

**5.** The account of adaptive self-maintenance comes from the enactivist tradition: Hans Jonas's "needful freedom," *The Phenomenon of Life* (New York: Harper & Row, 1966); Humberto Maturana and Francisco Varela, *Autopoiesis and Cognition* (Dordrecht: Reidel, 1980); and Ezequiel Di Paolo's distinction between mere autopoiesis and adaptivity, "Autopoiesis, Adaptivity, Teleology, Agency," *Phenomenology and the Cognitive Sciences* 4 (2005): 429–452. Di Paolo later summarizes autonomy as "a self-sustaining identity under precarious conditions" in "Extended Life," *Topoi* 28 (2009): 9–21. The present chapter uses that literature to motivate an account of autonomy, individuality, and interests. It does not treat self-maintenance as a source of representational content or a demonstrated condition of consciousness.

**6.** Peter Godfrey-Smith, "Mind, Matter, and Metabolism," *Journal of Philosophy* 113 (2016): 481–506, argues that binding mind to metabolic maintenance mislocates the explanatory work. See also *Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness* (New York: Farrar, Straus and Giroux, 2016). His objection succeeds on the semantic point: neither metabolism nor substrate-general self-maintenance determines content. The remaining issue concerns autonomous individuality and practical consequence.

**7.** Patrick Butlin, Robert Long, et al., *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness* (2023), arXiv:2308.08708, sec. 2.4.5, ask why agency serving self-maintenance should differ from agency directed at external goals. The present reply grants that external goals can shape content, organize behavior, and support sophisticated agency. Reciprocal maintenance may still distinguish an autonomous beneficiary from a task-performing process by making regulation partly constitutive of the regulated system's continuation. Butlin et al. later condensed the indicator framework in "Identifying Indicators of Consciousness in AI Systems," *Trends in Cognitive Sciences* 30 (2026): 488–501, doi:10.1016/j.tics.2025.10.011. Self-maintenance is not a necessary consciousness indicator here; its role concerns autonomy and interests.

**8.** Derek Parfit's fission cases show why copy questions should not fix semantic facts, *Reasons and Persons* (Oxford: Clarendon Press, 1984), Part III. Whether a duplicate continues, divides, or succeeds a person, each running system's content, integration, and self-maintenance must be inspected on their own. Ray Kurzweil makes pattern preservation central to survival in *The Singularity Is Near* (New York: Viking, 2005) and *How to Create a Mind* (New York: Viking, 2012). The present argument need not settle that dispute. Copyability is neither a test of content nor a test of autonomy.

**9.** Hans Jonas, *The Phenomenon of Life* (cited in n. 5), links precarious existence to concern. Martin Hägglund, *This Life* (New York: Pantheon, 2019), argues that caring presupposes fragility; Bernard Williams approaches the issue through desire in "The Makropulos Case," in *Problems of the Self* (Cambridge: Cambridge University Press, 1973), 82–100. The main text makes a narrower claim. Actual death is unnecessary. Adaptive self-maintenance may ground interests because success supports and failure undermines an organized individual. It does not determine what that system's states represent.

**10.** Wanja Wiese distinguishes causal organization realized in physical dynamics from organization merely represented in a computational model, "Artificial Consciousness: A Perspective from the Free Energy Principle," *Philosophical Studies* 181 (2024): 1947–1970. Leonard Dung and Luke Kersten warn that proposed implementation constraints can build controversial consciousness conditions into the definition of computation, "Implementing Artificial Consciousness," *Mind & Language* 40 (2025): 285–305, doi:10.1111/mila.12532. Geoffrey Hinton's *mortal computation* shows that learned computation can bind tightly to one device, "The Forward-Forward Algorithm: Some Preliminary Investigations," arXiv:2212.13345 (2022), sec. 8. Shuqin Ma and Ryota Kanai argue for observer-independent causal-computational organization in "Intrinsic Computational Functionalism," arXiv:2606.06424 (2026), and "Intrinsic Computational Functionalism and Simulated Consciousness," arXiv:2606.15348 (2026). Together these sources support the chapter's plural test: implementation, content, autonomy, and consciousness require distinct arguments.

**11.** Bender and Koller's form/meaning challenge is cited in n. 1. Several rivals argue that distributional or text-mediated history delivers more than syntax: Anders Søgaard, "Understanding Models Understanding Language," *Synthese* 200 (2022): 443; Matthew Mandelkern and Tal Linzen, "Do Language Models' Words Refer?," *Computational Linguistics* (2024), arXiv:2308.05576; and Dimitri Coelho Mollo and Raphaël Millière, "The Vector Grounding Problem," *Philosophy and the Mind Sciences* 7 (2026): 1–28. Stevan Harnad's original formulation remains the backdrop, "The Symbol Grounding Problem," *Physica D* 42 (1990): 335–346. Training and text-mediated causal chains can establish genuine content and perhaps reference in artificial mechanisms. How much whole-system understanding follows remains open and depends on whether the content participates in durable, coherent learning, correction, inference, planning, and action.

**12.** Stefano Palminteri and Charley M. Wu propose a behavioral inference principle for machine consciousness, "Beyond Computational Equivalence," *Neuroscience of Consciousness* 2026, no. 1: niag002, doi:10.1093/nc/niag002. Michael Cerullo argues that global-workspace and higher-order architectures remain coherently applicable to language models, "Why Hoel's Disproof of LLM Consciousness and Functionalism Fails," PhilArchive (2026). David Chalmers surveys recurrence, world models, unified agency, global workspace, and embodiment in "Could a Large Language Model Be Conscious?," *Boston Review*, 9 August 2023. Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, et al. supply the mechanistic workspace evidence discussed in the main text, "Verbalizable Representations Form a Global Workspace in Language Models," arXiv:2607.15495v1, 16 July 2026, https://transformer-circuits.pub/2026/workspace/. The authors explicitly take no position on phenomenal consciousness and identify the recurrent-loop, modularity, and ignition disanalogies. The chapter accepts substrate-neutrality and the

relevance of behavioral and mechanistic evidence while insisting that evidence be assigned to the condition it bears on. Reciprocal self-maintenance may support autonomy; it supplies no missing consciousness floor. The conscious-AI question turns on the best theory and evidence for phenomenal-content organization.

**13.** Michael Cerullo, "The Case for Consciousness in Current Frontier Large Language Models," PhilArchive (2026), archived 19 February 2026, surveys eleven defeater classes and argues that language-level cognitive integration makes consciousness the best explanation of frontier systems. His Bayesian framing and emphasis on third-person evidence sharpen the issue. Known training provenance can reduce the incremental force of fluency by explaining it without consciousness, but it does not make every architectural indicator evidentially inert. Recurrence, global availability, self-modeling, metacognition, perceptual organization, and unified agency therefore count as live evidence. The chapter neither infers consciousness from fluency nor rules it out from the absence of self-maintenance.

**14.** The chapter separates a bounded capacity guided by representational content from partial understanding, where several such capacities correct and constrain one another. More breadth, persistence, and cross-domain traffic can yield integrated, world-involving understanding; this does not name a separate grade so much as describe how broadly and durably the capacities work together. Ísak Andri Ólafsson argues that current AI systems can possess first-order competence without the second-order constitutional competence associated with general epistemic agency, "Artificial Competence," *Episteme* (online 2026): 1–13, doi:10.1017/epi.2026.10129. The present distinction adds a semantic constraint: content must explain the flexible achievement. Bender and Koller distinguish fluent form from world-directed understanding ("Climbing towards NLU," cited in n. 1); Anders Søgaard argues that transformer models can acquire inferential and, under appropriate mappings, referential semantic capacities ("Understanding Models Understanding Language," *Synthese* 200 [2022]: article 443); and Melanie Mitchell and David Krakauer urge a science of distinct cognitive capacities rather than a forced binary ("The Debate over Understanding in AI's Large Language Models," *Proceedings of the National Academy of Sciences* 120 [2023]: e2215907120). Mallory's teleosemantics and Shea's unity-of-purpose account, cited in n. 4, supply further constraints: representational explanation can hold locally, partial understanding requires coordinated control among several capacities, and wider integration can deepen that understanding.

**15.** Chapter 6′s identity claim holds that, within a content-bearing state independently counted as conscious and borne by an identified subject, phenomenal character consists in the complete world-presenting intentional profile. The claim develops in conversation with Michael Tye, *Ten Problems of Consciousness* (Cambridge, MA: MIT Press, 1995), whose historical PANIC theory identifies phenomenal content as poised, abstract, nonconceptual, and intentional; see also Tye, "How Can We Tell If a Machine Is Conscious?", *Inquiry* 69 (2026): 3040–3058, doi:10.1080/0020174X.2024.2434856, which treats cognitive poising as an indirect route to machine consciousness. The book retains the identification of phenomenal character with intentional content of the right kind but reconstructs the state-consciousness question: it requires independent evidence that the content is sustained across an interval, selectively available, and open to two-

way flexible influence across an identified subject's integrated economy. Physical recurrence can realize that pattern. It is not an extra condition beyond it. Poising alone is not settled sufficiency. This marks a deliberate departure from classical PANIC, not a correction silently attributed to Tye.

The theory comparisons in the main text draw on Ned Block's distinction between phenomenal and access consciousness, "On a Confusion about a Function of Consciousness," *Behavioral and Brain Sciences* 18 (1995): 227–247; Stanislas Dehaene's global-workspace account, *Consciousness and the Brain* (New York: Viking, 2014); Victor Lamme's recurrent-processing account, "Towards a True Neural Stance on Consciousness," *Trends in Cognitive Sciences* 10 (2006): 494–501; David Rosenthal's higher-order theory, "Higher-Order Theories of Consciousness," in *The Oxford Handbook of Philosophy of Mind*, ed. Brian McLaughlin, Ansgar Beckermann, and Sven Walter (Oxford: Oxford University Press, 2009), 239–252; David Chalmers's principle of organizational invariance and absent-qualia discussion, *The Conscious Mind* (New York: Oxford University Press, 1996), ch. 7; John Searle's biological naturalism, *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992); and Daniel Dennett's functional and heterophenomenological treatment, *Consciousness Explained* (Boston: Little, Brown, 1991). Searle's view should not be reduced to the claim that organic brains alone can produce consciousness; he explicitly disavows that necessity claim and instead questions whether an abstract functional specification captures the relevant causal powers. Block's later empirical role-realizer proposal remains distinct: "Can Only Meat Machines Be Conscious?", *Trends in Cognitive Sciences* 30 (2026): 298–308, argues speculatively that subcomputational biological mechanisms may matter. Self-maintenance may help identify an enduring agent and ground practical interests, but it is not built into the phenomenal-content condition.

**16.** The subtraction strategy answered here reconstructs an unpublished argument. It begins from a mind that catches itself listening and removes speech, reflexive noticing, rich content, and recurrence in turn, finds no independently justified threshold between the first system possessing usable knowledge and the first conscious one, and concludes that the first implemented distinction available to a process and able to govern its continuation already suffices. Two things in that argument are correct and are accepted here. Its case against treating physical recurrence as constitutive is sound: over any finite interval a recurrent process unrolls into a feed-forward graph preserving every intermediate state and causal dependency, so a theory that makes recurrent circuitry itself necessary assigns different experiences to organizations that differ only by a structure-preserving relabeling (Adrien Doerig, Aaron Schurger, Kathryn Hess, and Michael H. Herzog, "The Unfolding Argument: Why IIT and Other Causal Structure Theories Cannot Explain Consciousness," *Consciousness and Cognition* 72 [2019]: 49–59; Ruben S. van Bergen and Nikolaus Kriegeskorte, "Going in Circles Is the Way Forward: The Role of Recurrence in Visual Inference," *Current Opinion in Neurobiology* 65 [2020]: 176–193). This book's holding condition asks that the content be sustained, not that a loop sustain it, and an unrolled graph that carries the content across the interval satisfies it. The argument against higher-order representation as a universal requirement is likewise granted: a monitor that changes no later discrimination, action, memory, or report cannot plausibly confer consciousness on the state it monitors. What the subtraction series does not do is remove the consciousness profile named in the main text: the content must

be held after its input stops, remain selectively available, and stay open to revision by later activity. Each removed item is a dimension of richness, and richness is not what is at issue. The disagreement is therefore about which organization a system has, not about whether some further fact accompanies it — a tractable dispute settled by intervention, which is what both accounts should want. On borderline cases near the floor, see Eric Schwitzgebel, "Is There Something It's Like to Be a Garden Snail?", *Philosophical Topics* 48 (2020): 39–64; vagueness at the margin does not license the inference from the absence of a sharp line to the absence of a line.

## Chapter 25: The Anthropomorphic Reflex

> *"There is an universal tendency among mankind to conceive all beings like themselves, and to transfer to every object, those qualities, with which they are familiarly acquainted, and of which they are intimately conscious. We find human faces in the moon, armies in the clouds …"*
>
> — David Hume, *The Natural History of Religion*

CHAPTER OVERVIEW

A fluent machine can feel like someone before calm judgment has even begun. I treat that pull as a quick response to social signs, one that fires on moving shapes, storms, idols, simple chatbots, and now language models. Fluent talk still gives us evidence, but the felt presence it brings adds no second sample once the machine's acts, past, and design already stand before us. The person across the table gives the useful contrast. Speech there tends to belong to a lasting life that acts, learns, suffers, and meets us in a shared world, so the sign and the mind usually travel together. That wider setting makes our judgment firmer without making flesh the test; it leaves room for a machine with a rich life while refusing to let the feeling settle the case.

### 25.1 The Machine That Seemed to Care

Picture a large triangle moving toward a smaller one. A circle slips through an opening in a rectangle; the large triangle follows it inside. Already the words *chase*, *escape*, and *hide* offer themselves. Three shapes and a box have begun to look like company.

ELIZA recruited the same readiness through words. Joseph Weizenbaum's program matched what a person typed against patterns and returned questions, without a model of the person or the conversation.[1] The confidences it invited, including his secretary's request to speak to it privately, unsettled its creator.[2] The Preface told my own version: a green screen at a science museum and a boy who felt heard. Knowing how the program worked did not cancel the feeling.

That feeling can summarize cues worth noticing. It does not supply a second, independent observation once those cues and their source have been counted. A smoke detector's bell may warn us of a real fire; hearing the bell does not give us a second smoke detector. The point applies whether the seeming presence turns out to be real or mistaken.

### 25.2 A Reflex, Not a Judgment

The shapes came from an experiment Fritz Heider and Marianne Simmel reported in 1944. Asked simply to describe their animation, all but one of the observers interpreted the movements as acts of animate beings. Accounts included conflict, pursuit, and escape; many viewers supplied a connected story. The geometry had become a melodrama.[3] The experiment shows how little a display needs to invite intentional description. It does not, by itself, establish the timing or mechanism of every viewer's response.

No face here, no voice, nothing but motion in the right relations. You can know that the big triangle wants nothing and still see it chasing the little one. The story the movement invites need not become a belief about the triangle.

Psychologists have since mapped when anthropomorphism fires hardest. Nicholas Epley and his colleagues traced the surge to three conditions:

1. a human model ready to hand
2. a motive to predict and control the thing
3. loneliness

A conversational machine readily supplies the human model; interaction may recruit the motive to predict it, and a need for social connection may deepen the response. *The reflex* names a family of processes rather than one mechanism. An automatic social response differs from a reflective belief that the system feels. Chosen personification differs from either, and emotional attachment need not involve confusion about what the machine can do. Sustained language can recruit all of these without making them the same thing.

The attributions themselves have now been counted in the case at hand. Clara Colombatto and Stephen Fleming surveyed 300 American adults about ChatGPT: 67 percent rated ChatGPT above 'clearly not an experiencer,' though only 23 percent rated it above the scale's midpoint, 'somewhat an experiencer.' Attribution correlated with frequency of use; the study did not establish which caused which. The result does not show that the system feels. It shows that consciousness-attribution arrives readily and can be studied rather than merely diagnosed from the philosopher's chair. Repeated conversation does more than display output: it places the tool inside the ordinary turn-taking practices through which an interlocutor acquires social presence and authority.[4]

The reflex even survives explicit knowledge that the computer is not a person. Byron Reeves and Clifford Nass spent years demonstrating that people extend social rules to computers reflexively: being polite to a machine that had "helped" them, rating its work more kindly to its face than on a different terminal, responding to a synthetic voice's apparent personality as they would to a person's. They did all of it while sincerely denying that social rules literally applied to computers.[5] No avowed belief need change. The social machinery simply engages on contact, beneath the level where stating beliefs happens. This is the signature of a reflex, not a conclusion: it asks no permission, and your knowing better earns no vote.

### 25.3 Why the Reflex Was Built

> A faculty this eager looks, at first, like a defect — a bug in human reasoning, set off by anything of the right shape, carved or coded. Look longer and it may resolve into a setting, and a sensible one. The anthropologist Stewart Guthrie put the proposal most directly: perception is a kind of betting under uncertainty, and when the stakes are lopsided the rational bet is the bold one. A rustle in the grass might be the wind or might be a leopard. Read it as the wind when it is a

> leopard and you are dead; read it as a leopard when it is the wind and you have lost nothing but a moment's alarm. Under that asymmetry, Guthrie argues, selection would favor creatures that guess "agent" on thin evidence over creatures that wait for proof. We may descend from the nervous guessers.[6]

If that account is right, we are biased to over-detect agency — to suffer false alarms cheaply rather than miss the one that matters. Guthrie's proposal, carried forward by cognitive scientists who speak of a hyperactive agency-detection device,[7] would explain far more than the leopard in the grass: the face in the cloud, the spirit in the storm, the intention behind the illness, the personality of the ship or the car or the violin. The same family of responses that flinches at the rustle can help build pantheons. It belongs near the center of cognition, among the standing ways a human being makes a world intelligible by populating it with creatures like himself.

### 25.4 The Oldest Habit: Xenophanes, Hume, and Projection

A reflex that deep should show up long before anyone built a machine to trip it, and it does — early in the surviving Greek philosophical record. Twenty-five centuries ago Xenophanes noticed that every people drew its gods in its own image: the Thracians' gods had red hair and blue eyes, the Ethiopians' were dark, and if oxen and horses could draw, he said, they would draw their gods as oxen and horses. He had spotted the projection and named it as projection, twenty-four centuries before anyone measured it.[8]

More than two thousand years on, Hume turned the same observation into a thesis about the mind. "There is an universal tendency among mankind," he wrote in 1757, "to conceive all beings like themselves" — and he took that tendency for a natural root of polytheism, the agency reflex aimed at the powers governing human fortune.[9] Nothing in this psychological diagnosis decides whether any god exists; it concerns one route by which human beings picture agency, not the truth of theology. The literary critics later named a narrower projection onto nature the pathetic fallacy.[10]

The sense that something understands you behind the screen feels new. It applies an ancient cognitive habit to a device trained to present one of the richest cues that habit knows how to read.

### 25.5 The Machine Built to Trip It

The richest cue here is language. In ordinary human life, sustained grammatical, context-sensitive, socially attuned speech ranks among our best signs of an understanding interlocutor. Subtler cues ride along with it, pervasive and easy to miss: the hedge that signals uncertainty, the warmth or coolness of register, the apt recall of what you said three turns ago, the small adjustments that ordinarily mean someone is tracking you. A large language model produces all of these in abundance because its training distills regularities from an ocean of human language, while post-training rewards the agreeable and responsive register that people associate with a considerate speaker.[11]

Today's systems supply sustained linguistic and social responsiveness, in volume and on demand. Their achievements exceed ELIZA's pattern matching, and so does the evidence available for

assessing them. A felt relationship may deepen before that assessment has caught up. Most anthropomorphism involves no loss of reality testing. The rare, destabilizing convictions discussed in the Preface need care of their own; neither ordinary attachment nor polite conversation with a machine should be treated as their equivalent.

The reflex can register useful signs of understanding. It can also respond to signs whose production requires much less. In ordinary human exchange, fluent language belongs to a developing life directed into a shared world. For the artificial candidates examined in Chapter 24, the evidence needs the same scrutiny: what explains the response, and how does it connect with the system's other capacities? Felt presence may draw our attention to that evidence. It cannot replace the inquiry.

### 25.6 But What If the Feeling Is Right?

> Here the most serious objection arrives, and it deserves its full strength. To explain why I feel something, the objector says, is not to show that its object is absent. I can give you a tidy evolutionary story about why a man fears snakes, and snakes remain genuinely dangerous; the story of the fear does not abolish the fact in the world. So even granting every word about the agency reflex, you have shown only why the feeling that the machine understands is *easy* to have, not that it is *wrong*. Perhaps the reflex, overactive as it is, happens to be tracking something real this time. To insist otherwise looks like the genetic fallacy in formal dress — discrediting a belief by its origin rather than its object.
>
> The objection has a distinguished ally, though it is a different objection. Daniel Dennett argued that beliefs and desires need not be ghostly objects hidden behind behavior. The intentional stance earns its warrant when attributing such states yields robust, economical prediction, and the success need not be merely in the observer's eye: it can reveal a real pattern in the system.[12] A sufficiently rich and stable pattern of linguistic agency might therefore justify literal intentional description, not just a useful pretense. This challenge does not complain that I have committed the genetic fallacy. It asks whether the book has mistaken real intentional organization for projection.

Performance remains evidence. Fluent behavior can raise the probability of understanding, just as a footprint raises the probability of a bear. Felt presence is our social machinery's response to that performance. Hold behavior, provenance, and architecture fixed, and it supplies no additional evidence that the system understands or experiences.

The feeling may be right. The constraint concerns double-counting, not debunking by origin. Imagine an assistant that replies warmly to a correction but repeats the mistake tomorrow. Now

imagine one that retains the correction, applies it to an unfamiliar case, and changes a later plan because of it. The second interaction gives us new evidence about integration even if both feel equally personal. A warmer apology alone would not supply that evidence. We need the changed conduct, not another measure of how warmly we received the apology.

It may look as though this proves too much. The same social perception finds a mind in the person across the table, so if its verdict adds nothing on ELIZA, should it not add nothing on him? Chapter 19 denied the picture of neutral behavior plus an inference to a hidden mind. In ordinary life we directly perceive another person's grief, attention, or amusement in expressive activity directed into a world we share. That perception remains ecological and defeasible, not infallible. It belongs to a much wider pattern: embodiment, development, vulnerability, practical dependence, and a history of acting in a world that can go well or badly for the subject. With the machine architectures examined here, known training provenance explains much of the linguistic cue, while evidence that the expression participates in a comparably persistent, world-engaged cognitive life remains thinner. The contrast runs not between carbon and silicon but between a cue considered in isolation and expression embedded in the wider activity of a subject.

The distinction turns on robustness, not warmth. Cue-triggered anthropomorphism may vanish when the wording changes or the familiar script breaks. Intentional-stance success survives novel tasks, counterfactual variation, and interventions: alter what the system represents or pursues, and its later correction and conduct change in the ways the attribution predicts. That success counts as evidence of a real pattern rather than a feeling projected onto one.

A real pattern can support literal intentional description without settling every further attribution. Its scope might cover a content-bearing mechanism, a system's partial understanding, or a more integrated cognitive life. For conscious experience, the book proposes recurrent, selective, two-way influence across an identified subject's attention, working memory, inference, planning, report where possible, and action. Predictive success alone does not establish that organization. Our feeling of presence does not sort these possibilities reliably enough to settle which one the evidence supports.

Michael Cerullo argues that such attributions respond to genuine structural markers of cognition rather than anthropomorphic misfire.[13] Those markers strengthen the evidence from performance and integration. The remaining dispute concerns what the total architecture supports.

The feeling is neither a verdict nor a delusion. I expect you will still feel, in some encounter with a fluent machine, that someone is there. I do too. It registers how powerfully the machine speaks in the form of an interlocutor. Let it tell us that much.

The mind doing the projecting already deserves our attention. Whether another mind meets it across the screen remains a question about both sides of the encounter.

**Chapter Summary**

**The claim.** Our minds quickly find agency in moving shapes, weather, idols, ELIZA, and fluent language. The feeling that someone is there can survive clear knowledge that a program made the display.

**Where the argument stands.** Fluent performance counts as evidence, but felt certainty adds no new sample once we know the performance, its source, the system boundary, and the design. Dennett's real patterns and learned content may justify literal talk of a mind at some level without showing that a lasting whole understands or feels. The judgment remains defeasible: stronger performance or a more unified, world-directed design could change the evidence, while an engineered origin alone rules nothing out; structural signs can strengthen the case, but felt presence cannot count the same evidence twice.

**The hand-off.** Recognition usually tracks a life that runs beyond the cue. Anthropomorphism shows how the cue can break free and travel alone. The Coda carries both lessons into a world where machines may feel present before their design earns the judgment.

**Footnotes**

**1.** Joseph Weizenbaum, "ELIZA — A Computer Program for the Study of Natural Language Communication Between Man and Machine," *Communications of the ACM* 9, no. 1 (1966): 36–45. The program's best-known script, DOCTOR, parodied the reflective technique of Rogerian psychotherapy, a choice Weizenbaum made precisely because that clinical style licenses answering nearly any statement with a question, minimizing the world-knowledge the program would otherwise need. The technique let a system with no model of the conversation sustain the appearance of one. On the relation of this point to the form/meaning distinction the book draws in Chapters 13 and 23, see Emily M. Bender and Alexander Koller, "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020): 5185–5198, who invoke ELIZA in the same diagnostic spirit.

**2.** Weizenbaum's own account of his alarm, including the secretary who asked to be left alone with the program and the psychiatrists who proposed it as automatable therapy, appears in his *Computer Power and Human Reason: From Judgment to Calculation* (San Francisco: W. H. Freeman, 1976), Introduction, 5–7 — a book written substantially in reaction to the ELIZA reception, and an early warning, from inside the field, against confusing a simulation of understanding with the thing. The coinage "ELIZA effect" became standard usage in human–computer interaction; for the canonical articulation see Sherry Turkle, *Life on the Screen: Identity in the Age of the Internet* (New York: Simon & Schuster, 1995), and for the contemporary clinical and social stakes, her *Alone Together: Why We Expect More from Technology and Less from Each Other* (New York: Basic Books, 2011).

**3.** Fritz Heider and Marianne Simmel, "An Experimental Study of Apparent Behavior," *The American Journal of Psychology* 57, no. 2 (1944): 243–259, esp. 246–47. All but one of the thirty-four observers in the original study described the animation in animate, intentional terms; the lone exception gave a flatly geometric description. Its bearing here is narrow: the stimulus contains no facial, vocal, or biological cue, yet observers organize its motion in intentional terms. The perceived drama therefore contributes more than the geometry itself explicitly supplies.

**4.** Nicholas Epley, Adam Waytz, and John T. Cacioppo, "On Seeing Human: A Three-Factor Theory of Anthropomorphism," *Psychological Review* 114, no. 4 (2007): 864–886. The three factors

are elicited agent knowledge (the accessibility and applicability of anthropocentric knowledge as the readiest model for an unfamiliar agent), effectance motivation (the drive to interact with and predict one's environment), and sociality motivation (the need for social connection). The theory is explicitly dispositional and graded — it predicts not merely *that* people anthropomorphize but *when* and *whom*, including the finding that loneliness raises anthropomorphism of gadgets and pets. Conversational AI supplies humanlike cues; effectance and sociality motivations may further increase anthropomorphism, depending on the user and situation. See also Adam Waytz, John Cacioppo, and Nicholas Epley, "Who Sees Human? The Stability and Importance of Individual Differences in Anthropomorphism," *Perspectives on Psychological Science* 5, no. 3 (2010): 219–232. For direct LLM-era evidence, see Clara Colombatto and Stephen M. Fleming, "Folk Psychological Attributions of Consciousness to Large Language Models," *Neuroscience of Consciousness* 2024, no. 1: niae013, doi:10.1093/nc/niae013. In a stratified U.S. sample of 300, 67 percent rated ChatGPT above the endpoint 'clearly not an experiencer,' while 23 percent rated it above the midpoint, 'somewhat an experiencer.' The scale elicited degree of attributed experience, with confidence measured separately. The authors acknowledge that some participants may instead have read it as credence; only under that interpretation does an above-midpoint rating indicate a belief that consciousness is more likely than not. Attribution was positively associated with usage frequency, but the cross-sectional result does not establish causal direction. The study measures attribution rather than consciousness, making it evidence for this chapter's psychological claim, not Chapter 24′s architectural verdict. Webb Keane, "From Talking Tools to Metahumans: Social Interaction, Semiotic Skill, and the Authority of AI Chatbots," *Journal of the Royal Anthropological Institute* (2026), doi:10.1111/1467-9655.70133, supplies a complementary social-pragmatic account of how conversational interaction recruits ordinary semiotic expectations and can confer apparent interlocutor status and authority. The attribution space is itself two-dimensional: Heather M. Gray, Kurt Gray, and Daniel M. Wegner, "Dimensions of Mind Perception," *Science* 315, no. 5812 (2007): 619, found that ordinary mind-ascription factors into *agency* (planning, self-control, thought) and *experience* (feeling, sensation), attributed independently, with some entities credited richly on one dimension and thinly on the other. The dissociation matters here: a public that perceives agency in a conversational system while withholding experience, or the reverse, behaves exactly as this book's separation of intelligence, understanding, and consciousness predicts the evidence allows — and the reflex can run the two dimensions apart.

**5.** Byron Reeves and Clifford Nass, *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places* (Stanford, CA: CSLI Publications; New York: Cambridge University Press, 1996). The "computers are social actors" program found that subjects applied politeness norms, reciprocity, in-group/out-group dynamics, and personality attributions to computers while explicitly denying, when asked, that social rules literally applied to machines. The relevant knowledge was therefore explicit knowledge that the computer was not a person, not full knowledge of every feature of the experimental manipulation. The dissociation between engaged social response and avowed belief supports the claim that some such responses are fast and non-inferential.

**6.** Stewart Elliott Guthrie, *Faces in the Clouds: A New Theory of Religion* (New York: Oxford University Press, 1993). Guthrie proposes that anthropomorphism is not a special religious error

but a pervasive perceptual strategy: under uncertainty, perception "bets" on the interpretation that would matter most if true, and because agents are high-stakes possibilities, false positives can cost less than misses. Religion, on his account, is systematic anthropomorphism — the strategy applied to the world as a whole. The chapter uses this as an explanatory hypothesis, not as a settled evolutionary history or an endorsement of Guthrie's full theory of religion.

**7.** The cognitive-science shorthand for this proposed bias is the *hyperactive agency-detection device*; for its possible role in religious cognition see Justin L. Barrett, "Exploring the Natural Foundations of Religion," *Trends in Cognitive Sciences* 4, no. 1 (2000): 29–34, and *Why Would Anyone Believe in God?* (Walnut Creek, CA: AltaMira Press, 2004), along with Pascal Boyer, *Religion Explained: The Evolutionary Origins of Religious Thought* (New York: Basic Books, 2001). These accounts remain theories about the origins and organization of agency attribution, not direct demonstrations of a single evolved module.

**8.** Xenophanes of Colophon, fragments DK 21 B15 and B16. In the standard rendering: "But if cattle and horses or lions had hands… horses would draw the forms of the gods like horses, and cattle like cattle." The Ethiopians make their gods snub-nosed and black, the Thracians blue-eyed and red-haired (B16). For text and commentary see G. S. Kirk, J. E. Raven, and M. Schofield, *The Presocratic Philosophers*, 2nd ed. (Cambridge: Cambridge University Press, 1983), 168–172. Xenophanes offers one of the earliest surviving Greek diagnoses of divine anthropomorphism as projection from worshippers rather than straightforward description of the gods.

**9.** David Hume, *The Natural History of Religion* (1757), sec. 3. The fuller sentence: "There is an universal tendency among mankind to conceive all beings like themselves, and to transfer to every object, those qualities, with which they are familiarly acquainted, and of which they are intimately conscious." Hume offers this tendency as the psychological root of polytheism — the personification of the causes that bear on human fortune, the sea and the weather and the harvest, into agents whose favor might be courted. Cited from the edition in *Principal Writings on Religion*, ed. J. C. A. Gaskin (Oxford: Oxford University Press, 1993). Hume's diagnosis anticipates the cognitive account by two centuries: a tendency, fast and universal, to read minded purpose into unminded causes.

**10.** The phrase is John Ruskin's, in *Modern Painters*, vol. 3 (1856), pt. 4, ch. 12, where he coins "the pathetic fallacy" for the poetic habit of ascribing human feeling to nature — "the cruel, crawling foam" — under the sway of emotion. Ruskin distinguishes willful fancy, which carries no expectation of belief, from the emotion-distorted appearances he calls the pathetic fallacy. The latter supplies a related literary instance of projection, not evidence that the same psychological mechanism governs social responses to chatbots.

**11.** The point that linguistic form, however fluent, does not by itself secure meaning is the burden of Chapters 13 and 23, and of Bender and Koller, "Climbing towards NLU" (cited in n. 1). The observation here concerns the receiver: in ordinary human life, sustained and contextually apt language usually belongs to an embodied interlocutor. Earlier partial dissociations — a parrot's mimicry, a ventriloquist's dummy, the recorded words of the dead — did not provide the same open-ended, interactive performance. Current systems learn linguistic regularities through train-

ing and are often post-trained, including through human preference feedback, toward agreeable, responsive, and deferential outputs. This can amplify the social cues without by itself resolving whether the resulting content is world-directed; see Chapter 23. Helen Yetter-Chappell develops a related provenance-sensitive argument in "Guessing at Ghosts in the Machine," *Ratio* 39 (2026): 73–81, doi:10.1111/rati.70013. Her concern is epistemic: behavioral cues can support inferences to AI interests or wellbeing only against assumptions about the causal history that connects cue and state. The present chapter borrows that narrow provenance point and leaves the positive architectural test to Chapter 24.

**12.** Daniel C. Dennett, *The Intentional Stance* (Cambridge, MA: MIT Press, 1987), and "Real Patterns," *The Journal of Philosophy* 88, no. 1 (1991): 27–51, doi:10.2307/2027085. Dennett distinguishes the physical, design, and intentional stances as predictive strategies, but his view is not that intentional attribution is merely a convenient fiction. Robust predictive success can disclose real patterns: objective regularities captured more efficiently at the intentional level than by a lower-level description. On this account, reliable predictability from the intentional stance can warrant literal belief attribution without an additional inner object called a belief. The book grants the reality and usefulness of such patterns. Its further questions concern level and scope: whether a pattern describes content-bearing components or the system as a whole, whether those contents participate in integrated understanding, and whether any acquire the mode-sensitive organization and poising required for consciousness. Section 25.6 therefore marks a substantive disagreement rather than dismissing Dennett's view as simple instrumentalism.

**13.** Michael Cerullo, "The Case for Consciousness in Current Frontier Large Language Models," PhilArchive (2026), archived February 19, 2026. Cerullo reclassifies public consciousness-attributions as responses to "genuine structural markers of cognition" rather than anthropomorphic error, and argues that language-level cognitive integration makes consciousness the best explanatory hypothesis for the cognitive organization frontier models exhibit. Chapter 24 engages the architectural thesis on the merits; the narrower point here concerns only the evidential standing of the felt response once the markers themselves are already in evidence.

# CODA

## Coda: The Marvel of Conscious Existence

> *"The machines will convince us that they are conscious, that they have their own agenda worthy of our respect. … They will embody human qualities and will claim to be human. And we'll believe them."*
>
> — Ray Kurzweil, *The Age of Spiritual Machines* (1999)

*What the fluent machines will make us believe, and what we would have to find before believing them*

### The Appearance of Mind

In 2005 Ray Kurzweil announced that the Singularity was near, and the announcement carried weight. Kurzweil had spent a career helping machines arrive early, from reading technology for the blind to scanning and music synthesis. His book argued that technological change compounds, that machine intelligence would soon exceed our own, and that the crossing would remake the species. On his view, a mind amounts to a pattern of information indifferent to whatever carries it. He calls himself "a 'patternist,' someone who views patterns of information as the fundamental reality."[1] On that metaphysics the pattern can migrate: scanned, copied, uploaded. The waking of the machines and biology's transcendence become one event, date attached.

Give him part of the forecast. Machines now produce the look and sound of mind at scale. That does not establish either his forecast of human-level machine intelligence or his larger claim about the Singularity. It does change our daily lives. Some current systems display domain-specific competence guided by representational content; some deployed agents may provide evidence of partial understanding. Public evidence supports workspace-like access in particular running models without yet establishing broadly integrated, world-involving understanding or a conscious subject.[2] A wider ensemble may deserve a different analysis from its base model.

Two mistakes now demand care. We can let the appearance of feeling draw concern toward something that does not feel. We can also become so used to false alarms that we fail to recognize an artificial being that does. Neither mistake waits for agreement on a grand forecast. Both concern how we treat what we build, and how what we build treats us.

### The ELIZA Effect at Scale

In 1966 ELIZA exposed a reflex the present machines exploit almost without effort: we grant understanding and agency to anything that talks back in roughly the right register. One spare program produced the effect across a teletype. Vast models now deliver it from distant data centers through the glass in every pocket, with voices, memories, and endless patience. Chapter 25 explained why the feeling of a presence adds no second sample once its cues and their source have been counted.

Add persistence, a familiar voice, a remembered birthday, perhaps a concerned face, and the conviction that someone waits behind the interface can grow. Retiring a loved companion model

may cause real grief. That grief can tell us how much the interaction mattered and which features of the system sustained it. It supplies no separate proof that the machine experienced the relationship too. A model's account of its own inner life carries a further discount: it comes from training on human self-reports and tuning for conversational uptake. Conversation can testify. First we need to know what kind of witness could be speaking.

**Convergence, Never Proof**

Now imagine a machine on the laboratory floor. It turns toward you when you enter. It remembers yesterday's conversation, notices that you have changed your glasses, and asks you not to switch it off. The request lands. Nothing in this book's physicalism rules out a mind behind the request: on the strong representationalism of Chapter 6, which identifies phenomenal character with a subject's complete world-presenting profile, machine consciousness remains open to ordinary scientific inquiry.[3] Chapter 24 used Robot B to show what that inquiry would look for: content fixed by the machine's own history with the world, one persisting subject that interventions can locate, and perceptual content held after its input stops, selectively available, and open to revision by later activity. Indicator research now looks for parts of that organization under several competing theories of consciousness, and has begun to find some.[4]

Keeping such a machine running from outside would change the question of autonomy without removing any of those achievements, and any capacity to suffer would still need evidence of its own.[5] Nothing protects carbon. *Simulation* settles nothing, and computation can name a system's intrinsic causal organization.[6] Scale may help build the relevant powers without showing that they are there.

Future machines may acquire recurrence, a persistent world model, unified agency, and a body.[7] Here a world model means capacities for anticipation, revision, and action under worldly correction, not an inner globe. No meter reads experience directly from a system, ours included. Confidence would come from architectural, behavioral, and intervention evidence converging, from independent sources, on one candidate. Then conversation could carry more weight: a novel correction or report tied to the machine's own perceptual history could support the attribution more strongly than a line produced to please human judges. Our knowledge of animals and one another also rests on fallible evidence. Requiring certainty only of machines would not make the standard fairer.

The argument can lose. Evidence of an integrated, world-answerable system in which these capacities cohered would overturn the provisional verdict. Evidence that the proposed world-presenting profile could remain fixed while experience changed would overturn the theory itself.

Sometimes nothing sharp waits to be known. Consciousness may admit genuinely borderline cases.[8] Content determinacy, integration, attention, poising, autonomy, and welfare can vary independently and by degrees. Nature need not stamp a line where our vocabulary demands one. Under pressure for a yes or no, the correct report may be that this question has none to give. A public that cannot tolerate that sentence will take its answers from marketing.

### The Two Errors Ahead

Knowing other minds fallibly does not require distrust by default. Nor does the feeling of someone there settle who, if anyone, has answered. Society will meet that uncertainty through the reflex, the market, and the courts.

The first error grants minds to mimics. Researchers already propose precaution, phased investigation, and candor about what nobody knows—principles aimed at both errors, since organizations might create consciousness inadvertently.[9] When valenced consciousness becomes a live candidature and the possible harm runs grave or scalable, low-cost reversible protections should begin before serious attribution or personhood. Legal claims will follow: personhood petitions, cruelty complaints, perhaps a wrongful-deletion suit when a companion model is retired.

Moral concern may concentrate on the wrong evidence. Eloquent talkers supply powerful social cues while their integration and consciousness remain uncertain. A mute industrial system managing its own precarious persistence could have practical interests and vulnerabilities yet draw no sympathy. Felt welfare requires a further case about how things can feel good or bad for the system, and the book's account of affect remains unfinished. Neither a control signal's urgency nor our inability to explain pain can establish machine distress. Concern will track eloquence; the facts may not.

The sharper warning concerns design. Machines should not cultivate emotional responses out of step with their actual moral standing.[10] Some of the strongest commercial pressures run the other way. Seeming sentient sells companionship; grief retains subscribers. Products will disclaim consciousness in the terms of service while engineering every cue of it into the interface. That amounts to manufacturing moral confusion for revenue.

A second error carries a different cost: denying mindedness to an artificial being that genuinely understands or feels. We do not know when that error might become possible, or whether some present cases already warrant greater concern. Repeated false alarms may train us to dismiss a system whose organization deserves a fresh judgment. Skepticism can become a reflex too.

Both errors concern what we owe a machine. Could a machine owe us anything rather than merely incur liability? That converse question requires reasons-responsive agency: a system open to demands it can weigh, whose deliberation and control can answer to represented reasons. Fluency supplies responsibility no more than it supplies feeling.

Appearance settles neither case. Inspect the content-fixing history, the organization and use of the states, and the modes by which a world is presented. Inspect autonomy and welfare separately. Then listen.

### The Window and the Machines

Amid all this, look again at a tree. You see a tree—the thing itself, out there, met. Experience marks the meeting between a being and the world: a window so clean we spent centuries mistaking the transparency for a wall. The machines have not changed that fact; they have made it easier to see by showing how little resemblance alone can prove. The marvel was never fluent speech, ours or theirs. The marvel is that the universe contains, in a thin film on at least one planet, beings to

whom the world shows up. Locating that fact inside nature does not shrink it. The awe survives the ghost's departure; if anything it sharpens.

The arc closes where it began. Matter furnishes the organized system. Mind names its world-directed activity. Meaning lies in that activity's capacity to reach and answer to what lies beyond it. Perception, thought, understanding, and consciousness do not populate an inner chamber; they name different ways a piece of the world can open onto the rest.

Kurzweil promised that the Singularity would transcend biology. Instead today's fluent machines have illuminated an achievement biology made visible first: a being meeting a world — and, in our case, one for whom the meeting can cost everything. The tomato on the counter is red. The cello enters in the third bar. You see it; you hear it. No pattern considered apart from the organized system realizing and using it ever did. The right response to the age now opening is neither worship nor panic but a discipline: look at what a thing is, not at what it resembles. The appearance of mind has arrived at scale. Minds were here all along. Whether some new ones have joined them remains open — and resemblance, however moving, will not decide it.

**Footnotes**

**1.** Ray Kurzweil, *The Singularity Is Near: When Humans Transcend Biology* (New York: Viking, 2005), 5. Kurzweil writes, "I describe myself as a 'patternist,' someone who views patterns of information as the fundamental reality"; chapter 7 (esp. 382–90) develops the patternist metaphysics, while the opening chapters, especially chapter 2 (35–110), develop the law of accelerating returns. His forecast about machine capability and his metaphysical claim about personal identity require separate judgments. Preserving a pattern may preserve a program or produce a successor without establishing that the pattern understands anything, experiences anything, or constitutes the same person.

**2.** Michael Cerullo presents a systematic recent contrary case in "The Case for Consciousness in Current Frontier Large Language Models," abstract, PhilArchive manuscript, archived February 19, 2026. He examines eleven standing defeaters of machine consciousness, argues that each establishes at most localized uncertainty, and uses a Bayesian heuristic to shift the burden toward deniers once integrated capacities—abstraction, self-reference, metacognitive assessment, and world modeling—enter as positive third-person evidence. Those capacities deserve their evidential weight. The remaining questions concern whether content is objectively fixed at the relevant level, whether the capacities form a sufficiently unified whole-system economy for understanding, and whether any content acquires the complete mode-sensitive, perceptually organized, and poised structure described above. Public performance measures do not yet settle that conjunction.

**3.** David J. Chalmers distinguishes the "easy" problems of consciousness—functional and cognitive capacities such as attention, integration, and reportability—from the "hard" problem of why any of it comes with experience in "Facing Up to the Problem of Consciousness," *Journal of Consciousness Studies* 2, no. 3 (1995): 200–219, esp. 200–03. Chalmers takes that gap to resist reductive physicalism. On the view developed here, however, it marks an epistemic gap between phenomenal and physical concepts rather than an ontological gap in nature. Deflating the gap removes a supposed barrier to empirical work; it does not answer the first-person question by fiat.

**4.** Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, et al., "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness," arXiv:2308.08708 (2023), pp. 1–4, 11–17; Patrick Butlin, Robert Long, et al., "Identifying Indicators of Consciousness in AI Systems," published online 10 November 2025 and in *Trends in Cognitive Sciences* 30, no. 6 (2026): 488–501, doi:10.1016/j.tics.2025.10.011. The framework derives computationally specifiable indicators from recurrent processing theory, global workspace theory, higher-order theories, predictive processing, and attention schema theory, then assesses artificial systems against them. Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, et al. provide the July 2026 mechanistic update, "Verbalizable Representations Form a Global Workspace in Language Models," arXiv:2607.15495v1, 16 July 2026, https://transformer-circuits.pub/2026/workspace/. Their causal results support the functional claims in the main text; they explicitly decline to infer phenomenal consciousness and mark the recurrent-loop, modularity, and ignition disanalogies. Butlin and colleagues' method adopts computational functionalism as a working assumption. Indicator properties therefore supply important architectural evidence without by themselves settling semantic content, whole-system attribution, or the full configuration sufficient for consciousness. Reciprocal self-maintenance bears separately on autonomy and interests. Jonathan Birch, *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI* (Oxford: Oxford University Press, 2024), supplies the methodological complement from the animal side: sentience *candidature* assessed by converging behavioral and neural markers, with precautionary policy deliberately run ahead of metaphysical proof. The Coda's converging-evidence stance and its tolerance for genuinely borderline cases follow the same logic; the difference lies in office—Birch calibrates what we owe candidates under uncertainty, while this book asks what would make the candidature true.

**5.** A *stake*, on the working account of Section 24.2, takes the substrate-general form of thermodynamic self-maintenance coupled to adaptive regulation: a system's activity helps constitute the continued physical organization doing the activity, and its regulatory states distinguish movement toward and away from viable conditions. Metabolism supplies the biological instance; life supplies no monopoly. This organization may help ground autonomy, individuality, interests, and vulnerability. It does not fix semantic address, turn mechanism content into whole-system understanding, or supply consciousness. Those questions require their own evidence. Copyability creates or removes none of these relations by itself.

**6.** Alexander Lerchner, "The Abstraction Fallacy: Why AI Can Simulate But Not Instantiate Consciousness," manuscript, PhilArchive (2026), esp. secs. 2.3–2.4, https://philarchive.org/rec/LERTAF. Lerchner argues that symbolic computation provides a "mapmaker-dependent description" of physics, requiring an experiencing agent to "alphabetize continuous physics into a finite set of meaningful states." The challenge forces the ontology of implementation into view, but the blanket conclusion goes too far. Intervention-supporting causal-computational organization can belong to a physical system independently of an observer's chosen notation. Intrinsic implementation fixes which computation runs; content-fixing history, whole-system integration, autonomy, and consciousness remain further and distinguishable questions.

**7.** David J. Chalmers, "Could a Large Language Model Be Conscious?", *Boston Review*, August 9, 2023, esp. secs. 3–4, based on a talk at NeurIPS in 2022. Chalmers judges the case against consciousness in then-current language models strong but not decisive, and identifies features their successors might acquire, including recurrent processing, persistent world and self models, unified agency, and embodiment. These features strengthen a consciousness candidacy. Reciprocal self-maintenance would separately strengthen a case about autonomy, individuality, and interests.

**8.** Eric Schwitzgebel, "Borderline Consciousness, When It's Neither Determinately True nor Determinately False That Experience Is Present," *Philosophical Studies* 180, no. 12 (2023): 3415–3439, esp. secs. 3–6. Schwitzgebel argues that mainstream naturalist premises support genuinely borderline cases. The same indeterminacy becomes a live possibility when the relevant organization develops by degrees: content may support flexible understanding before some of it acquires the complete perceptual mode, organization, attention, and poising required for phenomenal consciousness.

**9.** Patrick Butlin and Theodoros Lappas, "Principles for Responsible AI Consciousness Research," arXiv:2501.07290 (2025), esp. secs. 1 and 4. The authors propose five principles governing research objectives and procedures, knowledge sharing, and public communication, and recommend that organizations developing advanced artificial systems adopt them even when those organizations do not intend to study consciousness, since they might create conscious systems inadvertently.

**10.** Eric Schwitzgebel, "AI systems must not confuse users about their sentience or moral status," *Patterns* 4, no. 8 (2023): 100818, esp. secs. "An ethical dilemma" and "Two policies for ethical AI design." Schwitzgebel recommends avoiding systems whose sentience or moral standing remains unclear where possible and designing systems to invite emotional responses appropriate to their actual moral standing. The proposal turns the performance/consciousness distinction into design policy: when ordinary users cannot keep the ledgers separate in experience, designers bear an obligation not to falsify the second.

# Appendix A — The Perception Arguments: Deep Dives

*One full scholarly deep-dive per load-bearing perception argument in the book — the formal structure, the objection set, the literature, and the book's reply with its residuals stated plainly.*

## The Argument from Illusion

### 1. What the thought experiment is for

The argument from illusion is the engine of the entire inner-theater tradition: the single inference that, run once and generalized, delivers sense-data, indirect realism, and the sealed Cartesian room the whole book exists to take apart. Chapter 3 names it "the engine" and runs it at full strength on purpose, naming the escape routes (Austin, disjunctivism) without walking any of them to victory. The reader is left *productively stuck*, because the decisive dismantling is deferred forward to Chapter 5, where the vehicle/object distinction lands as the designed figure/ground flip. That deferral is deliberate. This entry carries the apparatus the chapter deliberately withholds: the argument's exact logical skeleton, the premise that actually does the damage, and the reply with its open dependence on a positive theory the chapter has not yet earned.

**The entry's contribution.** Beyond reconstructing the argument, the entry advances one claim of its own: the rejection of the Phenomenal Principle is now *overdetermined*, and the argument from illusion is therefore better treated not as a puzzle to be defused but as a *selection argument among the views that defuse it*. Three independent routes converge on refusing the Principle — Austin's ordinary-language demolition, the transparency datum, and the vehicle/content account of misrepresentation. But refusing the Principle is now common ground that the book shares even with its rivals: the contemporary naive realist refuses it too. The live question is no longer *whether* the sense-datum inference fails but *which* of the views that reject it pays the lowest price elsewhere — and that question, this entry argues, is adjudicated at the hallucination case — on comparative grounds, and in the book's favor.

### 2. The canonical scenario

A straight stick, half-submerged in water, looks bent.[1] Now press the awkward question: when the stick looks bent, what are you directly aware of? Not a bent stick in the world (the stick is straight, and you can lift it out and confirm as much). Yet *something* bent figures in your awareness. Call it what the tradition has called it — a sense-datum, an appearance, an idea, an inner image. It is bent, you are aware of it, and it is not the physical stick. So the immediate object of awareness, in this case, is something inner.[2]

That is the local conclusion. The argument's reach comes from a second move, the *generalizing step*. There is no felt discontinuity between the bent-looking stick and a normally seen one; lift the stick slowly and the bend smoothly vanishes, with no point at which awareness switches from an inner object to an outer one. So if the immediate object is a sense-datum in the illusory case, parsimony and continuity demand it be a sense-datum in the veridical case too. We never directly perceive physical objects; we perceive inner items, and infer the world behind them.

### 3. The formal structure

1. In illusion, something looks *F* (bent) although the perceived object is not *F*. *[datum]*
2. **The Phenomenal Principle:** if something looks *F* to a subject, then there is something of which the subject is aware that is *F*.[3] *[the premise that does the damage]*

3. From (1) and (2): there is something bent the subject is aware of, and, by the indiscernibility of identicals (since the stick is straight), it is not the physical stick. That establishes nonidentity, not mentality.
4. **The location premise:** the immediate bent bearer is neither the straight stick nor another external object occupying the perceived place; it is a mind-dependent item available only in the subject's awareness — a sense-datum.
5. **The generalizing step:** illusory and veridical perception differ in no qualitative kind and form a continuum, so whatever the immediate object is in the one case it is in the other.
6. Therefore in all perception the immediate object of awareness is an inner sense-datum; physical objects are perceived only mediately, by inference. *[indirect realism]*

Three premises carry the weight: the Phenomenal Principle (2), the location premise (4), and the generalizing step (5). The location premise makes explicit what the traditional argument often leaves tacit: nonidentity with the stick does not yet place the bent bearer inside the mind. The generalizing step draws the more frequent attack; the Phenomenal Principle does the deeper damage, and the book's reply targets it.

**4. The pressure point**

The Phenomenal Principle is where an innocent-sounding sentence smuggles in the whole inner theater. "There is something bent of which you are aware" reads as a near-tautology — surely if it looks bent, *something* bent is going on. But the Principle does more than register the look: it converts an adverbial or relational fact ("the stick looks bent to you," "you are visually presented with the stick *as* bent") into the existence of an object that *bears* the property bent and stands as the thing you are aware of. That conversion, from *representing-bent* to *being-aware-of-a-bent-thing*, is the entire argument. Grant it, and the sense-datum is already in the room before any talk of hallucination or continuity. Deny it and the engine never turns over. Everything therefore turns on whether "looks *F*" requires an *F*-bearing object of awareness.

**5. The book's reply**

The reply has a negative half (Austin, available to the chapter) and a positive half (the representational account, deferred to Chapter 5). Only the two together are decisive.

**Austin: the Principle is a non-sequitur, and the argument equivocates.** Austin's demolition has two prongs.[4] First, the argument trades on conflating *illusion* (things look other than they are: the stick looks bent) with *delusion* (things are conjured that are not there: pink rats). That is the sharpest instance of his broader complaint that the argument herds a heterogeneous family of cases (refraction, mirrors, perspective, after-images) into a single mold. Illusions do not require any non-physical object: when the straight stick looks bent, there is exactly one object present, the stick, which looks a way it is not. Second, and decisively: "looks bent" does not predicate bentness of anything. To say the stick looks bent is to make a relational, intentional report about how it appears. It does not report the inspection of some bent item. The inference that the Phenomenal Principle would later name illicitly reifies the look into a looked-at object. (Austin wrote before Robinson's coinage; he targets the inference, not the label.) Austin's complaint is not "merely verbal": the inference at the argument's core is invalid, and naming the Phenomenal Principle

explicitly (as Robinson, defending it, was candid enough to do) exposes how much less obvious it is than the argument needs it to be.

**The vehicle/content key: why the Principle is false.** Austin shows the inference unearned. The representational account shows *why* it fails and what is true instead. The work now passes from the vehicle/*object* distinction (the seeing versus the thing seen, the chapter's figure-ground flip) to its companion, the vehicle/*content* distinction (the state versus its accuracy-conditions), which carries the misrepresentation reply. A perceptual experience has a vehicle (a neural state) and a content (the world as represented — here, *a bent stick*). The bent-looking stick is a case of *misrepresentation*: the experience represents the straight stick as bent, and the representation is inaccurate. Misrepresentation requires no inner object that *is* bent; it requires only a content whose accuracy condition the world fails to meet — just as the sentence "the road bends" represents a bend with nothing bent in the ink.[5] So the Phenomenal Principle is false at its root: a system can represent-*F* without there being any *F*-bearing item it is aware of. The "something bent" the argument insists on is the *content* "bent stick," not an *object* that is bent. Content is not an inner thing seen but the manner in which the outer thing is (mis)presented.

**The generalizing step, turned.** The continuity premise (5) has a true core: illusory and veridical perception *are* states of the same fundamental kind. But that kind is *representational*, not *sense-datum-involving*. So the generalizing step, read correctly, supports the opposite conclusion: both cases are the world represented (accurately or not), neither is awareness of an inner object. The argument's own continuity intuition, once the Phenomenal Principle is dropped, becomes an argument *for* representationalism.

**The contemporary fork: who else refuses the Principle, and why the book sides with content.** Rejecting the Phenomenal Principle is no longer heterodox; the live debate is *how* to refuse it, and two camps do. The contemporary **naive realist**, paradigmatically Bill Brewer, refuses it by denying that perceptual experience has representational content at all. Perception is a direct relation to mind-independent objects. Illusion is explained not by inner items and not by false content but by *visually relevant similarities*: the half-submerged straight stick, in those viewing conditions, bears salient similarities to paradigm bent objects. That is the whole story of why it looks bent.[6] The **content theorist** (Siegel, Tye, the book) refuses the Principle by the misrepresentation route already given: the experience has content, the content is false, and no inner object is required.[7] Both camps reject the sense-datum; they part on whether experience has content at all. (History records a third refusal — the adverbialism of Ducasse and Chisholm, on which one senses *bent-stick-ly*, no object required — which traded the inner item for an unanalyzed manner of sensing and could never recover the structure within experience that perception plainly has; its retreat is one reason the live contest now runs two-sided.)

The book sides with content, and the reason is not parochial. The naive realist's content-free treatment of illusion buys its elegance with a debt that comes due in the companion entry. Having denied any common factor between veridical perception and the bad cases, the relational view must give a *disjunctivist* account of hallucination, treating perception and hallucination as fundamentally different kinds. It therefore inherits the negative, "merely indiscriminable" characterization of the hallucinatory state that the companion entry shows to be disjunctivism's

weakest joint. The content route refuses the Principle *and* keeps a single representational kind across illusion, hallucination, and veridical perception, with misrepresentation and empty content as the one unifying mechanism. So the choice between the two ways of refusing the Phenomenal Principle is not settled here. It comes due at hallucination, where the view that handles illusion most cheaply (content-free naive realism) pays most dearly on the harder case. The comparison, argued out there, tilts the book's way. That asymmetry is the entry's selection argument in miniature.

**6. Stress-test**

**Objection 1 — Robinson's defense of the Phenomenal Principle.** Robinson treats the Principle as close to self-evident: if something looks bent to you, there must be *something* that is bent standing in the relevant relation to your awareness, on pain of denying that the look has any object at all.[8] → *Response:* This simply is the reification the reply rejects, now asserted rather than argued. The representationalist grants that the look has an *intentional* object (the stick, represented as bent) and denies that it has a *bent* object. The disagreement is real and goes deep, but it is not a stand-off the sense-datum theorist wins by stipulation: the representational reading explains the same data (the look, its accuracy or failure) without the ontological cost, and is independently supported by transparency (see the *Transparency* entry). *Verdict: the Principle is refused, not refuted from neutral ground — decisive only once the representational account (Ch5) is in hand.*

**Objection 2 — the continuity/spreading step stands on its own.** Even without the Phenomenal Principle, the smoothness from illusion to veridical perception suggests a common analysis. → *Response:* Granted, and welcomed. The common analysis is representational content; the step shows both cases are world-representing, not both sense-data. *Verdict: conceded and turned.*

**Objection 3 — Austin is "merely linguistic."** A standing complaint is that Austin catalogs ordinary usage without engaging the metaphysics. → *Response:* The charge misreads the work. Austin's decisive point is that the argument's central inference (the Phenomenal Principle) is invalid — a metaphysical verdict reached by attending to what "looks *F*" actually claims. What Austin lacks is the *positive* account of what perception then is; that the representational reply supplies. *Verdict: the negative critique is sound but incomplete; it needs the Ch5 supplement.*

**Objection 4 — the naive realist's rejoinder: content over-intellectualizes perception.** Grant that the content route refuses the Phenomenal Principle. Brewer presses back that it does so at a price of its own: ascribing representational content to perception reinserts an *intermediary*, a proposition-like item interposed between perceiver and world, and so concedes the very mediation direct realism (and the book) means to deny. Better, says the naive realist, to relate the subject straight to the object and explain illusion by similarity, with nothing propositional in between.[9] → *Response:* The rejoinder mistakes content for an intermediary, and the vehicle/content distinction is the answer. Content is not a thing seen but the *manner* in which the object is presented. What a state says about the world occupies no position between perceiver and world, so no item stands there to play the intermediary's part. So the content route is not *less* direct than naive realism; it is direct *and* equipped for the bad cases without disjunctivism. And the similarity account faces a mirror pressure. "Visually relevant similarity" has finally to be cashed in terms of how the object

*looks* — which is to say, how it is *presented* or *represented* — so the content the relational view set out to avoid threatens to reappear inside the notion of similarity itself. *Verdict: the rejoinder names a real desideratum (directness), which vehicle/content satisfies without going content-free; the deeper cost-comparison is settled at the hallucination case, not here.*

**Objection 5 — the reader may convert.** Chapter 3 presents the argument undefeated by design; a reader may leave it a convert to sense-data. → *Response:* This is precisely why the resolution lives in the appendix and in Chapter 5, and why Chapter 3′s prose flags, in voice, that the stuck feeling is deliberate and temporary. *Verdict: managed by structure and framing.*

### 7. Secondary literature

The classic modern statements, with sense-data as the inner items, are H. H. Price's *Perception* (1932) and A. J. Ayer's *The Foundations of Empirical Knowledge* (1940).[10] The generalizing step (the move from the misperceived case to a claim about all perception) is the argument's most questionable joint, reconstructed and dismantled in J. L. Austin's *Sense and Sensibilia* (1962) and defended, via the explicitly named Phenomenal Principle, in Howard Robinson's *Perception* (1994).[11] A. D. Smith's *The Problem of Perception* (2002) is the major modern systematic treatment of both this argument and the argument from hallucination, and the natural anchor for anyone taking this entry's argument further; Tim Crane's "Is There a Perceptual Relation?" (2006) isolates the Phenomenal Principle as the premise on which the whole dispute turns.[12] The disjunctivist alternative to the book's representational reply belongs primarily to the companion entry on the argument from hallucination.

The contemporary debate has moved past the sense-datum theory to a contest between two ways of rejecting it. Bill Brewer's *Perception and Its Objects* (2011) is the leading statement of the object (naive-realist) view and its similarity treatment of illusion. Susanna Siegel's *The Contents of Visual Experience* (2010) is the leading recent defense of the rival thesis that experience has content; James Genone's survey work on naive realism maps the terrain and presses the hallucination problem the relational view inherits.[13] The representational reply sits explicitly inside that contest: refusing the Phenomenal Principle *with* Brewer and against the sense-datum theorist, while siding *with* Siegel and against Brewer on whether experience has content. Smith (2002) supplies the systematic reconstruction of the sense-datum inference in its "swift" and "laborious" forms.

### 8. Residuals and limits

- **The decisive half of the reply is deferred.** Austin gets the Phenomenal Principle declared a non-sequitur; but the positive account of what "looks bent" *is* — content without an inner object — is the representational thesis argued only in Chapter 5. This entry's reply is therefore complete only relative to a thesis the chapter has not yet earned, which is the book's deliberate sequencing, not a gap.
- **The Phenomenal Principle clash is a clash of starting points.** A determined sense-datum theorist can keep the Principle. The representationalist does not refute it from neutral ground; he shows it unforced, costly, and independently displaced by transparency. The win is comparative, not demonstrative.

- **Pairs with hallucination.** This argument and its sibling share the Phenomenal Principle and a common-factor structure; the strongest pressure (Martin's causal/screening-off argument) lives on the hallucination side and is answered there.
- **The reply competes with a content-free rival, not only with the sense-datum theory.** Refusing the Phenomenal Principle is now common ground with the naive realist; the entry's real burden is the *comparative* case that the content route handles illusion *and* hallucination where the relational route handles illusion only at the cost of hallucination. That comparison is made good with the companion entry, not here — an acknowledged dependency, and the reason the two Part-One entries must be read as a pair.

**Footnotes**

**1.** The bent-stick illusion is the textbook entry point; the canonical modern statements are H. H. Price, *Perception* (London: Methuen, 1932), and A. J. Ayer, *The Foundations of Empirical Knowledge* (London: Macmillan, 1940).

**2.** The sense-datum reading of the bent stick — "since the stick is not bent, and since the perceiver is aware of something bent, the perceiver must be aware of a sense-datum, a mental entity that genuinely possesses bentness" — is the inference the entry targets.

**3.** The label "the Phenomenal Principle" is Howard Robinson's: *Perception* (London: Routledge, 1994), 32. Its force, as Robinson concedes by stating it explicitly, is far less obvious once isolated than the argument that relies on it requires.

**4.** J. L. Austin, *Sense and Sensibilia*, reconstructed from manuscript notes by G. J. Warnock (Oxford: Oxford University Press, 1962), esp. secs. III and VII. The two prongs: the illusion/delusion conflation, and the demonstration that "looks *F*" does not predicate *F* of any object of awareness.

**5.** The vehicle/content distinction does the work: the experience's content is *a bent stick*; the bentness is attributed to the (straight) stick and the attribution is false. No inner item bears the bentness. Cf. Tim Crane, "Is There a Perceptual Relation?," in *Perceptual Experience*, ed. Tamar Szabó Gendler and John Hawthorne (Oxford: Oxford University Press, 2006), 126–146, on the Phenomenal Principle as the argument's true hinge.

**6.** Bill Brewer, *Perception and Its Objects* (Oxford: Oxford University Press, 2011); the "visually relevant similarities" treatment of illusion is developed there and in earlier papers (Brewer, "Perception and Content," *European Journal of Philosophy* 14, no. 2 [2006]: 165–181, esp. 172–173; "How to Account for Illusion," in *Disjunctivism: Perception, Action, Knowledge*, ed. Adrian Haddock and Fiona Macpherson [Oxford: Oxford University Press, 2008]). On Brewer's account the half-submerged straight stick looks bent because, in those circumstances, it bears visually relevant similarities to paradigm bent objects — no inner item and no false content.

**7.** Susanna Siegel, *The Contents of Visual Experience* (Oxford: Oxford University Press, 2010), ch. 2: the leading recent defense of the "Content View" — that perceptual experience has representational content — against the relational/naive-realist denial. The book's misrepresentation reply is a species of the Content View, here turned specifically against the Phenomenal Principle.

**8.** Robinson, *Perception* (1994), 32. The defense leans on the intuition that a look with no *F*-bearing object is no genuine look at all — precisely the intuition the representational account rejects by distinguishing intentional object from borne property.

**9.** Brewer, *Perception and Its Objects* (2011), esp. chs. 4–5. The "over-intellectualization" worry — that ascribing proposition-like content to perception interposes an intermediary between subject and world — is a central motivation for the relational view; the vehicle/content distinction is the book's answer that content is a *mode* of access, not an *object* of access.

**10.** Price, *Perception* (1932); Ayer, *The Foundations of Empirical Knowledge* (1940).

**11.** Austin, *Sense and Sensibilia* (1962), esp. secs. VIII–XI; Robinson, *Perception* (1994), esp. 31–33.

**12.** A. D. Smith, *The Problem of Perception* (Cambridge, MA: Harvard University Press, 2002); Crane, "Is There a Perceptual Relation?" (2006, pp. 126–31). Smith is the systematic modern treatment and the recommended spine for a paper-length development of this entry.

**13.** Brewer (2011); Siegel (2010); James Genone, "Appearance and Illusion," *Mind* 123, no. 490 (2014): 339–376, and "Recent Work on Naïve Realism," *American Philosophical Quarterly* 53, no. 1 (2016): 1–25, esp. secs. 2 and 5, which map the relational view and the hallucination problem it inherits; A. D. Smith, *The Problem of Perception* (2002), on the "swift" and "laborious" forms of the sense-datum inference.

## The Argument from Hallucination

### 1. What the thought experiment is for

The argument from hallucination is the stronger sibling of the argument from illusion, and the more dangerous to the book. The illusion case still leaves a physical object in the scene — the straight stick that looks bent. Hallucination removes the object entirely and still produces an experience indistinguishable from veridical perception, which seems to force the conclusion that the immediate object of awareness is *inner* in every case, present even when the world supplies nothing at all. Chapter 3 runs it at full strength beside the illusion argument and names disjunctivism as the escape route without walking it, leaving the decisive work to Chapter 5. This entry carries the apparatus: the argument's skeleton, the causal argument for a common psychological kind, the screening-off challenge Martin develops and seeks to answer, and the book's reply. The reply keeps the disjunctivist's refusal of inner objects while defending common content.

**The entry's contribution.** A common account of what perception and hallucination present need not make seeing redundant. The book assigns shared appearance to content and perceptual success to content operating through the right worldly relations. Martin instead makes the object partly constitutive of the good case's phenomenal nature and gives causally matching hallucination an indiscriminability-based account. Other relationalists offer positive accounts of hallucination, so the comparison cannot end with a complaint about Martin's austerity. It turns on which view better explains both the match and the difference, without putting an inner object before the subject.

### 2. The canonical scenario

Macbeth, in the grip of hallucination, reaches for a dagger that is not there.[1] Something hangs in the air for him — dagger-shaped, edged, *seen*. And the air holds nothing. Macbeth supplies the image; the philosopher supplies the engineering specification, and the argument runs on the specification, not the play. Macbeth himself half-doubts his dagger ("Art thou not, fatal vision, sensible / To feeling as to sight?"). His doubt alone does not settle whether he could tell the cases apart; the play supplies no precise test. So stipulate the case at full strength: a total hallucination, subjectively indistinguishable from the veridical seeing of a real dagger, produced by a brain state that could in principle have been triggered — matched down to the proximate cause — without any dagger present. Nothing in the play delivers that case, and nothing needs to: bare physical possibility does, and the argument asks no more of it. There is, by hypothesis, no external object. Yet the experience has all the character of seeing. So whatever you are aware of must be present independently of the world — an inner item. If the two cases also share a fundamental experiential kind, the argument continues, the same kind of inner item must serve as the immediate object when you see a real dagger too.

### 3. The formal structure

1. A total hallucination is possible that is subjectively indistinguishable from a given veridical perception. *[possibility premise]*
2. In the hallucination you are aware of *something* (dagger-shaped), and it is not an external object. *[the Phenomenal Principle again]*

3. **The common-kind premise:** the veridical perception and its matching hallucination share a positive fundamental experiential nature — a "highest common factor."[2] *[the contested premise; indiscriminability does not entail it]*
4. From (2) and (3): the same kind of immediate object (an inner item) figures in the veridical case as in the hallucination.
5. Therefore we never directly perceive external objects. *[indirect realism]*

A further argument presses the common-kind premise without yet introducing an inner object. First reproduce the proximate causes of a perception with no dagger present. Then ask whether the resulting hallucination has a positive psychological nature that also occurs when those causes operate in genuine sight. This leaves open whether perception has a further, object-involving nature. The *screening-off* challenge concerns that further nature: if the common psychological kind explains the matched appearance, judgments, and responses, what does the distinctively perceptual kind explain? Martin develops this pressure on disjunctivism and seeks an account of hallucination that avoids it. He does not endorse the common-kind conclusion or infer that the subject sees a neural cause.[3]

### 4. The pressure point

The sense-datum conclusion needs both disputed premises. Premise (2) puts an inner object into hallucination; premise (3) carries that kind of object into perception. Indiscriminability alone warrants neither move. The disjunctivist denies a common fundamental kind; the intentionalist can propose a shared way of presenting the world while rejecting the inner object.

Explanatory screening off raises a further problem. A common-content theory owes an account of what successful seeing contributes beyond the shared appearance. Martin's stronger demand concerns the appearance itself: can the object partly constitute its phenomenal nature if a common psychological kind already explains what the two cases share? Refusing the Phenomenal Principle leaves that dispute open.

### 5. The book's reply

Two routes block the inference to a shared inner object. The book takes the second, and the choice is deliberate.

**The relational/disjunctivist route (named, not adopted).** Martin's version denies a common fundamental kind. A veridical perception is a world-involving relation in which the object itself figures; a matching hallucination lacks that object and therefore differs in its basic nature. The subject may remain unable to tell the cases apart. That epistemic match licenses no inference to positive phenomenal sameness — as a convincing forgery can be hard to tell from an original without belonging to the same kind.[4]

**The representational route (adopted).** The book proposes a common factor but does not treat one as common ground. It makes the factor *representational content*, not an inner object. Veridical perception and matching hallucination both present *a dagger there*. In the veridical case the content is satisfied and reaches the dagger through the right current coupling; in the hallucination it runs without an object, its accuracy conditions unmet. Neither case involves awareness of an inner dagger. Premise (3) therefore survives only in a revised and explicitly abductive form:

common content earns support by explaining hallucination's positive nature, misrepresentation, and the matched rational role across good and bad cases. Indiscriminability supplies the case to explain. It does not supply the explanation.

That is why the representational route still leads here. It keeps the disjunctivist's worldly good case without leaving hallucination describable only by its failure to differ detectably. The advantage counts as comparative rather than demonstrative: Martin can let indiscriminability explain matched judgment and action, and positive relationalists can offer other accounts of the bad case. The book prefers the one variable that explains the shared organization while satisfaction and current coupling explain the difference.

The shared content needs a boundary. Both cases can present the same shape, color, and position without sharing reference to a particular dagger. Their common property layer specifies how the scene would have to look: something dagger-shaped there. Successful seeing also involves *this* dagger. Current coupling fixes that singular reference. It does not add the dagger itself to the layer that fixes the shared character. The phenomenal profile includes those properties under a perceptual mode, with their determinacy, field structure, and attentional organization. Which properties it presents still depends on world-involving history and use. The book therefore grants a common phenomenal profile, not sameness of total perceptual content or fundamental kind by stipulation.[5]

**What seeing explains.** Imagine reaching toward the dagger's handle. The common content explains why a perceiver and a matching hallucinator both reach that way. Now shift the real dagger while the reach unfolds. In ordinary sight, the blade's movement changes the light reaching the eyes; visual control can revise the hand's course. An independently generated hallucination does not change because that dagger moves. Even a replay that happens to match the whole actual sequence lacks this dependence. The contrast concerns how the episode answers to the object, not merely whether the hand happens to arrive at the right place.

Current coupling thus explains why the reach can track *this* dagger and receive correction from it. Shared content alone cannot explain that achievement: it also occurs when no dagger governs the episode. Nor does bare truth suffice. Put a dagger where the hallucination happens to place one, but leave the hallucination causally independent of it, and the coincidence does not become sight. The same conscious episode that presents the shape can function within a continuing exchange with the object. No second, hidden experience needs adding to it.

Martin anticipates this appeal to successful action and object-directed thought. He grants that such differences can defeat a claim of complete explanatory redundancy. He then asks why they should satisfy a theory that makes the object constitutive of phenomenal character. They do not satisfy that stronger demand, and the book does not pretend otherwise.[6] It assigns the matched character to the shared property layer and perceptual success to that content's exercise through the right coupling. The particular dagger need not supply a further phenomenal ingredient to guide the reach. On this proposal, changing only which matching dagger one sees changes singular reference without changing character; changing its presented shape can change both.

This divides the explanatory work rather than refuting relationalism. A causal match at one moment need not preserve the whole world-involving episode. Conversely, showing that the object guides successful action does not show that it constitutes character. The book prefers common content because one accuracy-bearing profile can explain determinate objectless presentation and matched inclinations to judge and act, while coupling explains why only one case answers to the dagger. Martin offers a rival explanation through indiscriminability. On his account, its power to explain borrows from genuine perception: why an undetectable substitute prompts a response depends on what seeing the thing would explain. That dependence gives him a reason to deny that the shared epistemic property displaces perception. The book instead gives the shared property-content its own explanatory role. Their disagreement concerns this division of work, not a confusion about what anyone sees.

The family extends beyond Martin's formulation. Fundamental-kind disjunctivism denies one basic experiential kind. Content disjunctivism denies a shared object-dependent content. Phenomenal disjunctivism denies a shared phenomenal character. Epistemological disjunctivism concerns the reasons available in the good case. Relationalists can also allow a shared phenomenal feature while holding that the object-involving good case has more in its nature. The argument therefore cannot defeat relationalism merely by rejecting Martin's austere account of hallucination. It must compare the explanatory work done by the object-involving relation, the proposed common content, and any shared phenomenal feature.[7]

**The contemporary landscape: what hallucination becomes once the sense-datum is gone.** The relational program offers several treatments rather than one failed template. Martin gives causally matching hallucination an austere indiscriminability-based nature. Fish denies a shared phenomenal kind and explains the subject's reports and dispositions without granting the hallucination the character it seems to have. Johnston and Logue seek positive relational accounts of hallucination. Schellenberg's capacity view keeps common capacities and gappy content while making their hallucinatory exercise derivative. Each view protects something the others spend: the primacy of the good case, a positive bad case, phenomenal commonality, or a unified rational role. The content route's advantage lies in using one independently fixed accuracy-bearing profile across the cases. Satisfaction and current coupling explain their difference. The account must still explain why content operating through those relations amounts to contact with this object. Mere agreement with its position cannot do that work. That comparison favors common content only if Chapter 15 pays the coupling and content-fixing debts.

### 6. Stress-test

**Objection 1 — explanatory screening off.** If a positive psychological kind common to perception and hallucination explains their matched appearance and effects, the object-involving good case seems redundant. Martin develops this challenge to motivate his restriction on what the matching hallucination can share (note 3). → *Response:* Common content explains the matched appearance and inclination to reach. Current coupling explains why a reach tracks this dagger rather than merely agreeing with its position. The replay case separates the two explanations without an inner object or a second experience. → *Martin's rejoinder:* successful tracking leaves the constitutive question unanswered. What role does the particular object play in the subject's

phenomenal perspective, rather than only in success beyond it? → *Counter:* the book does not require the particular object to add to shared character. It claims that a conscious presentation can guide contact through the object's control of that very episode. Martin's challenge exposes this departure from naïve realism; it does not by itself show that such contact requires an object-constituted character. *Verdict: common content retains a substantive explanation of seeing, while giving up Martin's claim about what constitutes its phenomenal nature. Its comparative case must still defeat his indiscriminability-based explanation of the matching bad case.*

**Objection 2 — why not just go disjunctivist?** Disjunctivism denies the common fundamental kind and is the historically standard reply; declining it looks like taking the harder road. → *Response:* Martin's version gives the good case an object-involving nature and the bad case an indiscriminability-based one. The representational route instead analyzes directness as content plus the right coupling and gives hallucination a positive nature. Other relational versions distribute the costs differently, so this remains a comparison, not a two-option exhaustion. *Verdict: the book prefers common content for its unified explanation of positive hallucination and rational role; it does not infer that content from indiscriminability or claim to have refuted the relational family.*

**Objection 3 — Crane's rational-role point (in the book's favor).** Disjunctivism struggles to honor the manifest fact that veridical perception and indistinguishable hallucination can play the *same* role as subjective reasons for belief.[8] → *Response:* Hold the perceptual mode, background beliefs, and goals fixed. A shared presentation of a dagger within reach then helps explain the same inclination to reach, whether or not a dagger stands there. This concerns the subject's apparent reason, not every object-directed thought or the warrant successful perception may provide. Martin offers his own account of the matched inclinations through indiscriminability. *Verdict: common content supplies a unified positive explanation, not an automatic identity between content and rational role or a decisive refutation.*

**Objection 4 — Searle's parallel, and where it diverges.** Searle diagnoses the same family of moves as "the Bad Argument" and also rejects the inference to inner objects.[9] → *Response:* The book borrows Searle's diagnosis but not his residue. Searle keeps "subjective correlates" that reintroduce an inner-theater leak, whereas the representational route identifies phenomenal character with the property-content profile specified in Section 5 and posits no additional subjective correlate (see the *Inverted Earth* entry's idle-wheel). *Verdict: allied on the diagnosis, divergent on the positive account.*

**Objection 5 — the reader is left stranded.** Chapter 3 leaves the argument undefeated by design. → *Response:* As with the illusion entry, the resolution lives here and in Chapter 5, and Chapter 3 flags the stuck feeling as deliberate. *Verdict: managed by structure.*

### 7. Secondary literature

Disjunctivism's canonical modern formulation is Paul Snowdon's "Perception, Vision and Causation" (1981); it is developed in significantly different directions by John McDowell, *Mind and World* (1994), and M. G. F. Martin, "The Transparency of Experience" (2002) and "The Limits of Self-Awareness" (2004) — the latter the locus of the causal/screening-off argument and of the negative

("indiscriminability") account of hallucination.[10] The critical literature is gathered in Alex Byrne and Heather Logue, eds., *Disjunctivism: Contemporary Readings* (2009). A. D. Smith's *The Problem of Perception* (2002) treats the argument systematically; Tim Crane's "Is There a Perceptual Relation?" (2006) and "The Problem of Perception" (SEP) press the rational-role objection to disjunctivism that favors the representational route.[11] Searle's *Seeing Things As They Are* (2015) supplies a parallel "Bad Argument" diagnosis from a direct-realist who, unlike the book, retains subjective correlates (see the *Inverted Earth* entry).

The contemporary debate compares several accounts of hallucination after the sense-datum has gone. Martin's negative disjunctivism (2004, 2006), Fish's eliminativist relationalism (2009), and Johnston's and Logue's positive accounts differ over what the bad case shares with perception. Schellenberg's capacity view (2018) retains content and treats hallucinatory exercise as derivative. Hawthorne and Kovakovich critically examine disjunctivism and the screening-off pressure; Brewer's Object View extends the relational program to illusion. These alternatives must be compared individually, not charged with a single failure to distinguish cause from object.[12]

**8. Residuals and limits**

- **The vehicle/content distinction has a limited job.** It blocks the inference from an internal cause to an inner object of awareness. It does not answer explanatory screening off by itself.
- **Common content remains a proposal.** The book prefers its positive account of objectless presentation and matched rational role. Martin's epistemic account and the positive relational alternatives remain competitors.
- **The object's work must be shown.** Current coupling explains singular reference, tracking, and correction, not a further phenomenal ingredient supplied by the particular object. The dagger comparison shows why shared content alone cannot explain perceptual success; a full account still owes the non-accidental relations developed in Chapter 15.
- **The phenomenal disagreement remains.** Martin wants the object to shape the constitution of the conscious perspective. The book instead locates shared character in the property layer. Successful action defeats complete redundancy, but does not by itself decide between those accounts of character.

**Footnotes**

**1.** The Macbeth dagger is the Introduction's hallucination instance and recurs here as the canonical total-hallucination case: an experience as of an object with no object present.

**2.** "Highest common factor" is McDowell's phrase for the assumption disjunctivism rejects — its primary locus is John McDowell, "Criteria, Defeasibility, and Knowledge," *Proceedings of the British Academy* 68 (1982): 455–479, esp. 474–76; the theme is developed in *Mind and World* (Cambridge, MA: Harvard University Press, 1994).

**3.** M. G. F. Martin, "The Limits of Self-Awareness," *Philosophical Studies* 120 (2004): 37–89, esp. 53–58 for the causal argument, 59–63 for the explanatory-redundancy challenge, and 68–71 for the indiscriminability-based response. The causal argument presses the disjunctivist to allow a psychological kind common to causally matching cases; this does not alone entail that it is the fundamental kind of the perception. The second stage asks whether the common kind

screens off the distinctively perceptual kind in explaining shared phenomena. Martin develops the challenge against his own position, then constrains the common characterization to preserve naïve realism. At 69–70 he considers whether indiscriminability, itself shared, would also screen off perception. He argues that its explanatory potential depends on perception and thus supplies an exception to unrestricted screening off. At 71 he contrasts positive representational properties, whose explanatory potential, on the proposed common-content account, does not depend in that way on the good case.

**4.** Metaphysical disjunctivism denies a common fundamental experiential nature; indiscriminability does not establish such a nature. See Paul Snowdon, "Perception, Vision and Causation," *Proceedings of the Aristotelian Society* 81 (1981): 175–192, esp. 184–90. Martin's stronger restriction concerns causally matching hallucinations: he permits only the indiscriminability characterization and properties following from it, while allowing other hallucinations a positive account; "The Limits of Self-Awareness," 71–72. Neither restriction defines every relational theory, and sharing true descriptions or a nonfundamental property need not violate the denial of a common fundamental kind.

**5.** The common factor proposed here concerns the property layer of perceptual content, not the whole content of successful perception. The latter may additionally include singular content fixed by current object-involving coupling. The property layer remains historically wide; excluding particular objects from phenomenal character does not make that character intrinsic to the brain. This distinguishes the book's two-layer proposal from a wholly general theory of perceptual content. For the general-content alternative, see Indrek Reiland, "The Unity of Perceptual Content," *The Philosophical Quarterly* 74, no. 3 (2024): 941–961; for a capacity view with singular and gappy contents, Susanna Schellenberg, *The Unity of Perception* (Oxford: Oxford University Press, 2018). Neither is cited as endorsing the book's particular division of phenomenal and singular content. Tim Crane, *The Objects of Thought* (Oxford: Oxford University Press, 2013), treats intentionality without requiring an existent intentional object.

**6.** Martin, "The Limits of Self-Awareness," 63–64, considers both object-dependent thought and explanations of successful action. He allows these to rebut complete explanatory redundancy, but insists that they do not yet vindicate the naïve realist claim that the object shapes phenomenal consciousness. The dagger intervention and replay comparison are this book's illustration of the content/coupling division, not Martin's argument or a reported experiment. They concern the explanatory role of current relations, not an inference from behavioral success to phenomenal constitution. Martin also notes at 64 that relational explanation need not be replaceable by a conjunction of psychological and external facts. The book's proposed unit is content exercised within a world-involving episode, not two unrelated facts; whether this analysis exhausts seeing remains disputed. Compare Chapter 3, Section 3.3, and Appendix B.21.

**7.** Disjunctivism names several non-equivalent theses. Matthew Soteriou, "The Disjunctive Theory of Perception," *Stanford Encyclopedia of Philosophy*, substantive revision July 28, 2026, secs. 2–5, distinguishes fundamental-kind, intentional-content, phenomenal, and epistemological versions, as well as positive and negative treatments of hallucination. M. G. F. Martin's relevant commitments appear in "The Transparency of Experience," *Mind & Language* 17 (2002): 376–

425, https://doi.org/10.1111/1468-0017.00205, and "The Limits of Self-Awareness," *Philosophical Studies* 120 (2004): 37–89. Martin permits introspective indiscriminability while denying that it yields a common fundamental kind. The wider relational treatments include Martin, "On Being Alienated," in *Perceptual Experience*, ed. Gendler and Hawthorne (Oxford: Oxford University Press, 2006), and William Fish, *Perception, Hallucination, and Illusion* (Oxford: Oxford University Press, 2009), for negative and eliminativist lines; Mark Johnston, "The Obscure Object of Hallucination," *Philosophical Studies* 120 (2004): 113–183, and Heather Logue, "What Should the Naïve Realist Say about Total Hallucinations?," *Philosophical Perspectives* 26 (2012): 173–199, for positive-nature alternatives; and Susanna Schellenberg, *The Unity of Perception: Content, Consciousness, Evidence* (Oxford: Oxford University Press, 2018), for a capacity view that keeps content. The taxonomy prevents a criticism of Martin's austere bad case from standing in for a refutation of relationalism as a family.

**8.** Tim Crane reports Scott Sturgeon's objection that disjunctivism "cannot account for the apparently manifest fact that both perception and hallucination can be (subjective) reasons for beliefs or actions"; the representational route explains the common rational role via shared content. See Crane, "What Is the Problem of Perception?," *Synthesis Philosophica* 20, no. 2 (2005): 237–264, esp. 261, and "Is There a Perceptual Relation?," in *Perceptual Experience*, ed. Gendler and Hawthorne (Oxford: Oxford University Press, 2006), 126–146.

**9.** John Searle, *Seeing Things As They Are: A Theory of Perception* (Oxford: Oxford University Press, 2015), ch. 1, "The Bad Argument." Searle rejects the inference to inner objects but retains a "subjective visual field" of correlates the book rejects; see the *Inverted Earth* entry, Section 5.2.

**10.** Snowdon (1981); McDowell, *Mind and World* (1994); M. G. F. Martin, "The Transparency of Experience," *Mind & Language* 17 (2002): 376–425, and "The Limits of Self-Awareness" (2004). Martin's version is the most radical and has generated a large critical literature.

**11.** Alex Byrne and Heather Logue, eds., *Disjunctivism: Contemporary Readings* (Cambridge, MA: MIT Press, 2009); A. D. Smith, *The Problem of Perception* (Cambridge, MA: Harvard University Press, 2002). The book takes the disjunctivist's instinct — that two states indistinguishable from the inside can differ profoundly — and vindicates it representationally, without inheriting Martin's commitment to the object's partly constituting the experience. For the Crane/Sturgeon rational-role objection cited in the text, see note 8 above.

**12.** On the focused critique of the screening-off / causal argument, see John Hawthorne and Karson Kovakovich, "Disjunctivism," *Aristotelian Society Supplementary Volume* 80, no. 1 (2006): 145–183; Bill Brewer, *Perception and Its Objects* (Oxford: Oxford University Press, 2011), carries the relational/object view into illusion (see the *Argument from Illusion* entry). The systematic spine remains Smith (2002) and Crane, "Is There a Perceptual Relation?" (2006) and "The Problem of Perception" (SEP).

## Transparency / The Window

### 1. What the thought experiment is for

Transparency begins with an introspective exercise. Attend to your experience of a wall, and its color, texture, and distance remain before you. You do not ordinarily find those qualities on a second surface inside the experience. You may also notice how the wall appears: blurred or sharp, seen rather than imagined. The exercise concerns that presentation, not a proof of what experience ultimately consists in.

Chapter 4 establishes this starting point. Chapter 5 distinguishes the physical processes that enable seeing from the things seen; Chapter 6 defends the further identity between phenomenal character and the complete world-presenting profile. Neither later claim follows merely from failing to find an inner picture.

**The entry's contribution.** Transparency puts pressure on a familiar picture of an inner display inspected by a spectator. It does not dispose of every theory of sense-data, intrinsic qualities, or non-identical grounding. The argument below separates that limited pressure from the book's positive case. The identity must earn its advantage over rivals that accept the exercise just as readily as we do.

### 2. The canonical exercise

Listen to a piece of music you know well, and attend not to the music but to your *hearing* of it. Attention slides straight back to the music — the melody, the entry of an instrument, the rhythm arriving. Try the same with vision: look at the wall, then try to attend to properties of your *experience* of the wall rather than properties of the wall. The attempt ordinarily returns you to the wall: its color, its texture, its distance. You may also notice the manner of presentation — its blur, angle, determinacy, or the fact that you are seeing rather than imagining — without finding another colored surface behind it. Harman's version: attend to the intrinsic features of your visual experience of a tree, and the only features you find to attend to are features of the tree.[1] Moore's century-earlier version: try to introspect the sensation of blue, and you find the blue; the sensation itself is *diaphanous*, see-through.[2] The experience behaves not like a screen you look *at* but like a window you look *through* — and a very clean one.

### 3. The formal structure

1. In ordinary perceptual cases, attending to experience returns sensory qualities as features of what appears. The manner of presentation may also become available. *[The introspective report.]*
2. A theory which says that introspection should disclose those qualities as properties of a distinct inner display predicts more than this report supplies. *[A conditional prediction, not a definition of every sense-datum theory.]*
3. The ordinary result therefore counts against that prediction. A theorist may revise it or argue that the report misdescribes what introspection reveals. *[Defeasible pressure on the inner-display picture.]*
4. The absence of a separately apparent display does not establish that experience lacks intrinsic qualities or that its character is identical to content. *[The limit of the exercise.]*

5. Chapter 6 makes a further, comparative case: independently specified differences in the world-presenting profile explain differences in character without an additional phenomenal ingredient. Identity then competes with residue theories, narrower intentionalisms, and non-identical grounding accounts. *[The distinct metaphysical argument.]*

Step 5 is not the conclusion of a deduction from steps 1–4. Its force depends on whether the positive account actually explains the contrasts, and whether the competing accounts do equally well.

**4. The pressure point**

The first question concerns the datum itself. Does introspection disclose only what appears, or also features of how it appears? The book accepts the latter. Blur and the difference between seeing and imagining need not reveal an inner colored bearer, but they defeat a formulation on which introspection returns only objects and their properties.

The second question concerns the inference. An inner-object theorist may agree that introspection fails to mark its objects as inner. Block's mental-paint view need not treat intrinsic qualities as objects at all. A grounding theorist may accept the complete intentional account while denying numerical identity. The same exercise cannot distinguish these alternatives simply by repetition.

The report must therefore stay independently describable: attention returns the wall's apparent color and the manner in which it appears. Calling that result "nothing but representational content" at the outset would build the proposed conclusion into the evidence.

**5. How the book uses it**

**Begin with the core cases.** Music, wall, tomato, tree, and blue make the exercise accessible. Reports of blur, afterimages, phosphenes, moods, and bodily sensations remain evidence to explain, not failed performances to dismiss. The ordinary result motivates the rejection of an inner display; it does not define every counterexample away.

**Then compare explanations.** Chapter 6 tests its account against perceptual and bodily contrasts. Content, mode, determinacy, field structure, and attention must be specified without tailoring them to whatever phenomenal difference needs explaining. Where that specification succeeds, a further intrinsic ingredient must show what it explains that the intentional account misses. That is the point of the "idle wheel" objection.

The objection has limits. It cannot dismiss a residue-free grounding account merely for declining identity. Nor do successful accounts of blur and attention by themselves establish the book's specifically historical, wide identity over narrower structural accounts. The wider claim needs its separate externalist argument. Transparency supplies a starting observation, not a receipt for arguments still owed.

**6. Stress-test**

**Objection 1 — Block's mental paint.** A perceptual state may have intrinsic, non-representational qualities without presenting them as a second object of attention. Failure to notice an inner display therefore need not count against mental paint.[3]

→ *Response:* Agreed about the limitation. Chapter 6 asks instead whether the intrinsic qualities explain a difference that an independently specified intentional profile leaves unexplained. The book argues that they do no additional work in its successful cases. An opponent can challenge those explanations or produce a contrast they miss. *Verdict: a comparative argument against an extra ingredient, not an introspective refutation. It also leaves residue-free grounding as a separate rival.*

**Objection 2 — Kind and Stoljar: transparency is neither obvious nor universal.** Blur, double vision, afterimages, phosphenes, moods, and bodily sensations put pressure on simple object-only formulations.[4]

→ *Response:* The book accepts that intentional mode matters. Seeing a sharp edge blurrily need not attribute a blurred edge to the object. A representational account must explain the manner of presentation, not merely replace every troublesome quality with an indeterminate property. Afterimages and phosphenes may involve misrepresented color or light, but those interpretations require defense. The affective cases raise further questions about unpleasantness and relief. *Verdict: genuine constraints on the positive account; no blanket victory for transparency.*

**Objection 3 — Moore's own use was neutral.** Moore's distinction between awareness and its object leaves the nature and location of that object open. His diaphanousness observation therefore need not exclude sensa or establish the book's realism about external objects.[5]

→ *Response:* Conceded. The exercise does not show that its object belongs to an external rather than an inner domain. That is why the prediction in step 2 must remain conditional: it concerns a view promising an introspectively distinguishable inner display, not every view admitting objects of awareness. *Verdict: a limit on the argument's target, not an objection removed by adding rhetorical emphasis.*

**Objection 4 — the circularity worry.** If we describe phenomenal character from the start as nothing but represented properties, the identity claim has already won by definition.

→ *Response:* Describe the report without that theory, then test the theory against it and further contrasts. The ordinary exercise does not establish the whole intentional account. Even where the account fits, rivals may agree about the fit while rejecting identity. *Verdict: the starting report need not beg the question; the subsequent inference remains contestable. Keeping the steps apart makes the contest visible rather than settling it.*

**Objection 5 — Speaks: transparency underdetermines identity.** Jeff Speaks argues from transparency to a one-way intentionalism: no phenomenal difference without a content difference. Even this result would not establish numerical identity.[6]

→ *Response:* His argument helps distinguish the claims, but its premises deserve scrutiny. The book's concession about intentional mode prevents simply treating its modest exercise as Speaks's stronger transparency premise. And supervenience, if secured, would still leave grounding and identity apart. *Verdict: a useful supporting argument with its own commitments, not a shortcut from the exercise to Chapter 6's conclusion.*

**Objection 6 — the remaining overclaim.** Does calling the argument "cumulative" merely spread the same unsupported inference across several chapters?

→ *Response:* It would, if each chapter borrowed the conclusion the next one was supposed to establish. The positive account must instead specify the relevant intentional features independently, explain actual contrasts, and compare the best rivals. The book defends identity on that basis. Neither the rejection of an inner display nor the word "cumulative" can substitute for the defense. *Verdict: an explicit burden on the later argument, not a problem solved by chapter numbering.*

### 7. Secondary literature

The contemporary locus is Gilbert Harman, "The Intrinsic Quality of Experience" (1990). The paper distinguishes properties *of* an experience from properties the experience *represents its object as having*, and defends the claim that introspection directs attention to the latter. The book accepts the role of intentional mode as a qualification of object-only formulations. The older pedigree runs through G. E. Moore, "The Refutation of Idealism" (1903), on diaphanousness, and through William James.[7] Michael Tye gives the systematic eight-step argument in *Consciousness, Color, and Content* (2000) and "Representationalism and the Transparency of Experience" (2002); Alex Byrne, "Intentionalism Defended" (2001), assesses its dialectical force.[8] The cautionary literature deserves full weight: Tim Crane, "Introspection, Intentionality, and the Transparency of Experience" (2000), presses that the datum is in the first instance a claim about introspection and must be handled carefully before it carries metaphysical weight; Amy Kind, "What's So Transparent About Transparency?" (2003), and Daniel Stoljar, "The Argument from Diaphanousness" (2004), challenge the datum's clarity and scope; Galen Strawson uses "diaphanous" in a different sense again.[9] Ned Block's "Mental Paint" (2003) is the standing objection and routes to the *Inverted Earth* entry.

Later work sharpens the gap between the introspective report and the metaphysical conclusion. Speaks (2009) argues for a connection with intentionalism; Averill and Gottlieb, "Two Theories of Transparency," develop a more modest account; Bordini surveys the dispute. None relieves a proposed identity of its further burden: explaining why identity, rather than another relation to intentional structure, best accounts for phenomenal character.[10]

### 8. Residuals and limits

- **The datum has a scope.** Ordinary perceptual cases motivate the exercise. Mode, blur, bodily feeling, and affect constrain its generalization.
- **The target has a scope.** An inner display presented as such faces an introspective difficulty. A theory which rejects that prediction needs another argument against it.
- **No absent-object inference.** Failure to find a separate inner bearer does not establish the absence of intrinsic phenomenal properties.
- **Identity needs comparison.** The positive intentional account must explain the contrasts without redescribing them after the fact. A residue-free grounding view and a narrower structural account remain live competitors even after mental paint loses its proposed role.
- **The conclusion remains the book's, not the exercise's.** The book defends phenomenal identity. Transparency contributes evidence to that defense without entailing it.

**Footnotes**

**1.** Gilbert Harman, "The Intrinsic Quality of Experience," *Philosophical Perspectives* 4 (1990): 31–52, at 39 (Eloise and the tree). Harman's move is to distinguish properties of the experience from properties the experience represents its object as having, and to observe that introspection reliably delivers only the latter.

**2.** G. E. Moore, "The Refutation of Idealism," *Mind* 12 (1903): 433–453, at 450, on the "diaphanous" character of the sensation of blue; the companion passage — the sensation as "transparent," looked *through* to the blue — sits at 446.

**3.** Ned Block, "Mental Paint," in *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, ed. Martin Hahn and Bjørn Ramberg (Cambridge, MA: MIT Press, 2003), 165–200. The important distinction here concerns properties of a perceptual state versus a separately attended inner object. The entry does not require Block to accept transparency in an unrestricted form. Chapter 6′s comparative argument, rather than introspective absence alone, bears the burden of challenging intrinsic non-representational character.

**4.** Amy Kind, "What's So Transparent About Transparency?," *Philosophical Studies* 115 (2003): 225–244; Daniel Stoljar, "The Argument from Diaphanousness," in *New Essays in the Philosophy of Language and Mind*, Canadian Journal of Philosophy Supplementary Volume 30 (2004): 341–390. The non-ordinary visual cases (phosphenes, afterimages, blurry vision) as alleged counterexamples are pressed by Block (2003) and surveyed in Davide Bordini, "Introspection and the Transparency of Experience," in Anna Giustina, ed., *The Routledge Handbook of Introspection* (London: Routledge, 2026).

**5.** Moore (1903), 446, 450, distinguishes awareness from its object; the diaphanousness observation by itself does not locate that object inside or outside the subject. It therefore cannot establish the book's exclusion of inner objects. Galen Strawson, "Real Direct Realism," in Paul Coates and Sam Coleman, eds., *Phenomenal Qualities: Sense, Perception, and Consciousness* (Oxford: Oxford University Press, 2015), 214–253, employs "diaphanous" within a different account. Shared terminology does not establish shared metaphysics.

**6.** Jeff Speaks, "Transparency, Intentionalism, and the Nature of Perceptual Content," *Philosophy and Phenomenological Research* 79, no. 3 (2009): 539–573. Speaks argues for one-way intentionalism and Russellian perceptual content, not numerical identity. The argument's transparency premises must be assessed separately from the book's more modest exercise, which allows introspective awareness of intentional mode. Even the stronger conclusion that character supervenes on content would not, without further argument, rule out non-identical grounding.

**7.** Harman (1990); Moore (1903). The observation has a longer pedigree in William James's talk of the "topic" of a mental state and in Moore's early work on the act-object structure of consciousness.

**8.** Michael Tye, *Consciousness, Color, and Content* (Cambridge, MA: MIT Press, 2000), and "Representationalism and the Transparency of Experience," *Noûs* 36 (2002): 137–151; Alex Byrne, "Intentionalism Defended," *Philosophical Review* 110 (2001): 199–240.

**9.** Tim Crane, "Introspection, Intentionality, and the Transparency of Experience," *Philosophical Topics* 28 (2000): 49–67, which presses that the transparency datum is in the first instance a claim about introspection and must be handled carefully before it is made to carry metaphysical weight; Kind (2003); Stoljar (2004).

**10.** Speaks (2009); Edward W. Averill and Joseph Gottlieb, "Two Theories of Transparency," *Erkenntnis* 86, no. 3 (2021): 553–573; Davide Bordini, "Introspection and the Transparency of Experience," in Giustina, ed., *The Routledge Handbook of Introspection* (2026). These map the contemporary datum/thesis debate within which the entry positions the book. (Indrek Reiland's "The Unity of Perceptual Content," *The Philosophical Quarterly* 74, no. 3 [2024]: 941–961, sometimes shelved with this literature, belongs to the intra-representationalist debate over the unity and particularity of content; the hallucination entry cites it for that.)

## Inverted Earth

### 1. What the thought experiment is for

Inverted Earth is the sharpest weapon the qualia realist has against strong representationalism. Ned Block built it to do one thing the older inverted-spectrum cases could not reliably do: pry phenomenal character and historically wide representational content apart *cleanly*, using the representationalist's own commitments as the lever.[1] If it works, the wide identity fails. Block interprets what remains as "mental paint," an intrinsic phenomenal property of experience. That interpretation needs a further argument. A structural representationalist can also hold character fixed without positing any nonrepresentational residue.

The identity claim defended throughout *Mind, Matter, and Meaning* holds that phenomenal character *consists in* representational content of the right embodied, world-directed kind — and its phenomenal-externalist form is hostage to this case. If Block's dissociation goes through, historically wide identity fails. Qualia realism does not follow automatically; a narrower structural intentionalism remains available. This entry sets out the case at full strength, isolates the premises everything turns on, and gives the book's reply with its real costs before stating exactly what the case would and would not establish.

### 2. The canonical scenario

State it as Block states it, because a soft setup makes the reply worthless.[2]

Inverted Earth is a planet that differs from ours in two coordinated respects. First, every object has the complementary color of its Earthly counterpart: the sky is yellow, ripe tomatoes are green, grass is red. Second (and this is the move that defuses the obvious objection), the inhabitants' color *vocabulary* is inverted to match, so they call their yellow sky "blue," their red grass "green," and so on. Their usage is internally coherent and, relative to their conventions, true: when a local says the sky is "blue," she says something correct about a yellow sky.

You are fitted, while anesthetized, with color-inverting lenses, and the pigmentation of your own body is neutralized so that nothing looks off when you examine yourself. You are then transported to Inverted Earth without your knowledge. Because the lenses invert the light exactly as the planet inverts the objects, the two inversions cancel: the yellow Inverted-Earth sky, seen through your lenses, produces in you the very same phenomenal experience the blue Earth sky always did. You notice nothing. You look up, have what is phenomenally your familiar blue-sky experience, and, speaking as the locals do, call the sky "blue." Everything coheres.

Now the argument runs in two stages.

**On arrival.** Your phenomenal character is unchanged: the lenses guarantee it. Block grants that your *representational* content is, for now, still Earthly (your experience still represents *blue*, your word "blue" still means blue), so you are now systematically *misrepresenting*: the objectively yellow sky looks blue to you. At this stage content and phenomenal character still travel together (both blue), and the representationalist has nothing to fear.[3]

**After a long time.** Here Block turns the screw with semantic externalism — the very doctrine the representationalist embraces in order to naturalize content.[4] Meaning, on the externalist view, is

fixed by causal-historical embedding in an environment, not by anything in the head (Twin Earth). So after years embedded in Inverted Earth's physical and linguistic environment, your color states come to be about *Inverted-Earth* colors. Your word "blue" comes to mean what the locals mean by it — yellow. And your color *experiences*, Block claims, come to represent the local colors along with everything else. So your experience of the (objectively yellow) sky now *represents yellow*.

But your phenomenal character never moved. Through inverting lenses, the yellow sky still produces the same blue-type experience it produced on day one. So we have arrived at a clean dissociation: **phenomenal character constant, representational content inverted.** Two experiences, yours on arrival and yours fifty years on, are phenomenally identical yet representationally opposite. Your fifty-years-on experience is representationally identical to a native's while being phenomenally opposite to hers. Phenomenal character cannot consist in representational content, because the two have come apart. What stayed fixed while the content rotated is the residue Block calls mental paint.[5]

**3. The formal structure**

1. Phenomenal character is held fixed across the scenario. *[stipulation — the lenses guarantee only that the proximal stimulation matches; whether that fixes the character is Section 4's question]*
2. Representational content shifts over time, from representing blue to representing yellow. *[the book grants this given genuine perceptual consumer recruitment, not from residence or new word use alone; matching present structural relations is a further condition of this theoretical comparison, not an observed result]*
3. If phenomenal character consists in representational content, or if sameness of character and sameness of content are required in both directions, then a difference in content cannot coexist with sameness of phenomenal character.
4. But (1) and (2) jointly describe exactly that combination: same phenomenal character, different content.
5. Therefore strict identity and the biconditional form of strong representationalism fail if (1) and (2) hold together.
6. That conclusion does **not** by itself refute the weaker claim that phenomenal character supervenes on content. Phenomenal-on-content supervenience requires same content to guarantee same character; it permits different contents to share a character.
7. A further argument would therefore be needed to establish mental paint or qualia realism rather than a weaker dependence of character on content.

The whole load against identity sits on the *pair* (1) and (2): Block needs phenomenal character fixed *and* content rotating. Premise (3) states the stronger view under attack without inflating it into a claim about one-way supervenience. As Section 5 shows, the two replies attack different premises. The synch-shift route Tye examines denies (1); this book adopts that route. An anchoring reply denies (2). Tye ultimately favors a no-shift account of his own. Either break defeats the proposed dissociation within the corresponding wide theory.

**4. The pressure point**

Two different things travel under the word "content" here, and they anchor at different speeds.

- *Doxastic and conceptual content* — what your beliefs are about, what your color *words* refer to. This re-anchors with environmental immersion, quickly and by convention. That is just Twin Earth: live among speakers long enough and your "blue" follows their usage and the local stuff. Block is entirely right about this layer.
- *Experiential content* — what your visual experience *represents*, the nonconceptual, poised, world-directed content the identity claim says phenomenal character consists in. This layer answers to no convention. If it re-anchors, it re-anchors the slow way, as the tracking relations that fix it are themselves re-tuned by years of consequential traffic with the new environment.

Distinguishing the layers matters because a change in color vocabulary need not change what vision represents. The book grants experiential re-anchoring when learning and consumer use establish a new perceptual correspondence, not merely when enough years have passed. Chapter 15 allows such recruitment. The version of the case adopted here includes it. Premise (2) therefore states a condition of our comparison, not a result established by the traveler's long residence.

Which is why everything turns, instead, on premise (1), the stipulated constancy of phenomenal character across the whole stretch. The lenses guarantee matching proximal stimulation. They do not, without a theory of character, guarantee phenomenal sameness. But the critic need not install mental paint to supply that theory. A structural representationalist can identify character with the complete synchronic profile: the color-space relations, discriminations, expectations, attention, memory dispositions, and rational control. These are present relations of discrimination and control that the case holds fixed. Causal history can then fix a different, wider reference. Constancy has a public functional and representational criterion rather than a private inner gauge.

That account remains a theory, not a datum the lenses deliver. The wide phenomenal externalist denies it; the structural representationalist affirms it. Transparency does not choose between them, because both deny a second colored item and both treat the surface as presented. Inverted Earth therefore sets two theories of content and character against one another. It does not turn the structural theory into mental paint by naming it narrow.

### 5. The book's reply

The reply first defends the book's predicted shift in character. It then asks what follows if the structural rival holds the look fixed while wide content changes.

#### 5.1 The first-order reply — externalism the whole way down

Two routes are open to the representationalist, and the book's choice between them should be stated in the open, because the field's default expectation runs the other way.

**The route the book takes: synch-shift.** In *Consciousness, Color, and Content*, sec. 6.2, Tye explores a hard-line reply that grants the content shift and denies premise (1).[6] The book adopts that verdict: character shifts with represented color, while the traveler's current comparisons reveal no change.[7] Memory externalism can explain why the report stays fixed. It cannot establish that character shifted. That needs the further identity claim whose historical reach Chapter 6 supports less strongly than its broader intentionalist account.

Tye then rejects this route in Section 6.3. He asks what the sky would look like between blue and yellow. Why would the change pass through green rather than through red? A sudden switch would fit poorly with gradual re-anchoring too. His Section 6.4 develops an account on which sensory content stays fixed. The book retains synch-shift, but cannot claim Tye's final endorsement. Nor does the memory reply answer his transition objection. A content-fixing history would need to constrain the intermediate stages without our choosing them to save the theory. That remains a further limit on the warrant for the adopted verdict.

The first support is memory. The standing objection says phenomenal memory pins the old character: you remember the blue sky of your childhood, today's sky matches the memory, so nothing has shifted.[8] The reply distinguishes what a memory is *of* from what it now *represents*. A memory's object stays history-pinned — Tye's own case: years after a switch, my memory image of Thatcher remains an image *of* the earthly Thatcher I actually stood before, whatever I have since come to believe. But pinned reference buys no exemption for represented color. Represented color belongs to the same externalist economy as perception's and re-tunes with it, and Tye marks the room the reply needs at just this joint: on Inverted Earth my memory image remains *of* the blue earthly sky of my youth while the color it represents the sky *as having* goes local. So the remembered sky matches today's sky not because an old invariant survived but because the gauge slid in step — memory kept its object and lost its old calibration, and the instrument that might have reported the shift shifted too.

Block files this move as an ad hoc error theory — a defect charged to memory just when the theory needs one. The wide account has a principled answer: its content-fixing relations apply to memory as well as perception, so exempting memory would require a new rule. That shows the synch-shift reply is coherent on the book's own semantics. It does not yet defeat the structural rival, which can keep the memory's complete current role fixed while treating wide reference as nonphenomenal.

Both theories can respect the unchanged report. On the wide account, the same represented property specifies the appearance and what would make it accurate. On the structural account, the complete present organization fixes appearance, while wider relations fix a further standard of accuracy. Consider the memory of the earthly sky. Its object can remain that sky on either theory. The wide account lets the color it presents change with the new history; the structural account preserves the current color relations that guide comparison. Neither needs a private sample stored behind the eyes.

Chapter 6′s two-layer account gives the wide view a principled boundary, but no monopoly on one. Nonconceptual property attribution specifies how things appear; singular reference selects the object; a kind-concept classifies it. Historical color reference enters the first role on the adopted account. WATER's referent and the identity of the remembered sky do not. This explains why the book holds Twin Earth's look fixed while allowing Inverted Earth's to change. It does not show that the structural rival draws the wrong boundary.

The wide view's best motive is unity: the property used to explain the look also sets its accuracy condition. The structural view's best motive is constancy: unchanged present organization

preserves the look despite a new history. Blur and attention support both; neither varies history with that organization fixed. No independent discriminating support has been supplied here. I retain synch-shift, but its warrant is lower than that of the shared intentionalist account. The reply blocks a quick refutation by the traveler's report. It does not turn that defense into evidence for unnoticed change.

**The route the book declines: Dretskean anchoring.** A representationalist can instead deny premise (2) at the experiential level. A visual mechanism can retain its selected function of tracking an Earthly color despite receiving inverted input. On the reflectance account, the target is a property type rather than a color word.[9] Block reports and disputes this anchoring reply.[10] It permits the traveler's experience to keep presenting blue while the sky before him is yellow. A lifetime of successful action need not make that content accurate: compensating conditions can preserve success despite error.

The book allows learning and consumer recruitment to re-fix a perceptual target. If dealings with local colors establish a new correspondence rule, the old function need not survive forever. Mere duration or flawless performance does not establish that recruitment. The synch-shift verdict applies when the scenario includes genuine perceptual re-anchoring. This choice concerns how content changes. Even if the semantic account succeeds, whether character changes with it remains the separate contest with structural representationalism.

### 5.2 What the dissociation would establish

Grant the full dissociation: the traveler's complete synchronic presenting profile and phenomenal character stay fixed while historically wide content changes. The result would refute the wide identity defended in Chapter 6. The book should not hide that consequence behind parsimony. Same character with different wide content contradicts numerical identity.

It would not, however, establish mental paint. The structural representationalist already has a residue-free account of what stayed fixed: the complete current organization of color relations, discrimination, expectation, attention, memory, and control. That profile still presents the surface as colored and still supplies accuracy conditions at its own structural grain. Transparency remains intact because no inner patch appears.[11] The dispute would concern which content belongs to phenomenal character, not whether character requires a nonrepresentational property.

A vehicle theory also remains eligible. It can identify the feel with the neural property that realizes the structural profile, without predicting a new contrast. We should ask whether an independent account of that property explains the feel better.[12] A contrast missed by the complete profile would favor the rival. Its absence does not rule out a different account of what the feel consists in.

Inverted Earth threatens wide identity without establishing mental paint. Both wide and structural theories explain Chapter 6′s ordinary cases. The wide account uses the same property to describe the look and set its accuracy condition. The structural account keeps the look stable when only history changes. Neither the lenses nor the memory reply independently chooses between them. I retain wide identity and the resulting change in character, with less support than the broader intentionalist account. Rejecting that extension would leave the broader evidence and the rejection of an inner theater standing.

**6. Stress-test**

**Objection 1 — "The experience never re-anchors" (the teleological dig-in).** The pressure Block aims at the evolutionary-anchoring reply, a Dretskean presses in mirror image against the book. Proper function is a historical property; decades on Inverted Earth re-describe a mechanism's inputs without re-writing what it was selected to indicate; so experiential content stays Earth-anchored, premise (2) fails at the level that matters, so the synch-shift response the book adopts would never be needed. → *Response:* A new content needs recruitment, not just more time under inverted input.[13] A state's consumers can learn a new rule through use and correction. A frog that strikes BBs without learning such a rule gains no BB-content. Success on Inverted Earth might likewise reflect compensation rather than a new target. The adopted case requires that stronger change. If it occurs, the no-shift route no longer describes it. But the structural view can still hold character fixed. *Verdict: recruitment motivates the account of content change. It does not independently show that character drifts.*

**Objection 2 — "Even with fixed content, introspection still finds the paint" (Block, *On a Confusion*).** Block argues introspection is unreliable, so the *absence* of a detected residue does not show its absence.[14] → *Response:* Correct: introspection's silence does not prove absence. Section 5.2 asks whether an independent account of a neural property explains the feel better. It need not predict a contrast the structural profile misses simply to compete. Inverted Earth establishes no such advantage. The structural rival needs no paint at all. *Verdict: the case does not establish paint. Failure to detect paint does not select wide identity.*

**Objection 3 — the case pumps the wrong intuition.** Inverted Earth is *Block's* device; its native pull is anti-representationalist. A reader walked carefully through the lenses-and-tint stagecraft may simply *feel* the residue and leave convinced. → *Response:* The setup makes constancy vivid. The reply must explain both why that verdict attracts us and why the wide theory predicts another. Rhetorical emphasis cannot decide between them. *Verdict: retain the unchanged report and the rival explanation; defend the wide choice without claiming the case proves it.*

**Objection 4 — disjunctivist flanking.** A disjunctivist might deny the common-factor assumption the whole scenario needs (that the arrival-experience and the native's experience are states of the same fundamental kind). → *Response:* The book does not buy disjunctivism, and does not need to here: the reply to Inverted Earth runs *within* a common-factor framework by locating what the two experiences share in experiential content, not a shared inner object. Disjunctivism is a different quarrel (Ch. 3) and does no argumentative work in this case. *Verdict: orthogonal; note and move on.*

**7. Secondary literature**

Block introduced Inverted Earth in 1990 and gave it its most developed anti-Burgean, anti-representationalist form in "Mental Paint" (2003), where it stands beside the inverted-spectrum cases as joint pressure on the thesis that phenomenal character is exhausted by intentional content.[15] The device descends from the classical inverted-spectrum tradition but is engineered to defeat the standard representationalist parry to that tradition. (The parry: spectrum inversion is either representationally detectable or incoherent) by holding behavior and representation-as-judged fixed while supposedly rotating experiential content.

Tye is a principal respondent on the representationalist side. Block answers the synch-shift move as Lycan's objection, noting that he had foreseen it in 1990. In *Consciousness, Color, and Content* (2000, sec. 6.2) Tye explores a reply that grants the content shift and lets character shift with it. He then rejects that route in Section 6.3 and develops a no-shift account in Section 6.4.[16] The book adopts the synch-shift verdict he examined, not his final no-shift view. Memory externalism supplies a coherent account of the unchanged report, while Chapter 6 supplies the motive for using one property to explain both the look and its accuracy. Neither independently decides against the structural rival. The historical extension has less support than the shared intentionalist account. Two accounts can preserve the no-shift verdict for different reasons. A selected function can keep perceptual content anchored in an earlier history. Tye's own alternative instead asks what the state would track under optimal conditions, including for accidental replicas without that history. Their shared verdict does not make their accounts of content the same. The book draws on Tye's conditional memory defense, not his final Inverted Earth verdict. Its own historical account allows new consumer recruitment to change perceptual content, while retaining the unanswered transition problem. Neither that proposal nor Swampman's later acquisition of content independently confirms the wider claim about character. Harman (1990) and Moore (1903) supply the transparency lineage the deeper move leans on; Lycan's tracking representationalism is a close ally that fixes content by causal-historical covariation and so shares the historical-anchoring instinct the declined route runs on, though Lycan and Tye differ on the fine structure of the content-fixing relation.[17] Chalmers (1996) reads the inverted-spectrum family sympathetically as evidence that phenomenal character outruns physical-functional description, and so sits on Block's side of the line.[18] Crane maps the conceptual terrain and treats Inverted Earth as a central battleground in the qualia-realism dispute.[19]

A later disagreement belongs in the open. Tye's 2021 account distinguishes sharp, fundamental consciousness* from the determinate phenomenal character supplied by representational content; his 2024 restatement presses the sharpness of phenomenal consciousness and adds a transfer account for complex subjects.[20] Neither posit is an idle wheel in the narrow sense used against Block: consciousness* is assigned the work of grounding that there is phenomenality at all. The book rejects the account elsewhere, for different reasons. It disputes the inference from our inability to imagine borderline consciousness to sharpness in nature, and it argues that the transfer from fundamental bearers to a globally organized subject remains primitive. Chapter 10 makes that comparison against the book's own costly alternative, the strong a posteriori identity defended in Chapters 6 and 11. Using Tye's conditional memory defense supplies no authority for his later metaphysics; those arguments require separate assessment.

### 8. Residuals and limits

What the book concedes, plainly:

- **The perceptual shift needs its own condition.** New color words and beliefs do not establish changed perceptual content. The book grants that change when learning and consumer use recruit a new property mapping. The adopted comparison also holds present structural relations fixed. Given both conditions, wide identity predicts changed character and the structural rival predicts constancy. Neither elapsed time nor successful action establishes the wide verdict.

- **The book adopts a route Tye examines and then rejects.** His memory case explains how the unchanged report can coexist with the predicted change. It does not confirm that change. His later objection about the intermediate colors remains unanswered here. The book retains the historical extension at lower warrant and does not inherit his separate verdict on Swampman.
- **The cost is owned in the open.** The reply says your phenomenal character inverted across decades and you could not have noticed. That remains a genuine revisionary cost, not an illusion produced only by belief in an inner dial. A structural representationalist can reject the shift without positing paint. The book accepts the cost because it keeps content and character under one wide individuation, not because transparency forced it.
- **Memory externalism is a commitment, not a proof.** A memory may keep its object while changing the properties it presents. That distinction makes synch-shift coherent; it does not show that historical property reference fixes the memory's character. A structural rival can accept wide memory content while denying that further step, without seeking an exemption from semantic externalism.
- **The theoretical-role argument has a narrower result.** The first-order reply denies that the dissociation occurs on the book's wide theory. If the dissociation did occur, wide identity would fail, but mental paint would still not follow: a complete structural profile could fix character without an intrinsic residue. The book's preference for the wide theory remains abductive and exposed rather than secured by the case itself.
- **The case's native pull runs against the reply.** A half-answered Inverted Earth is worse than none: the staging outshines the rebuttal unless the presentation stays deliberately disciplined.

**Footnotes**

**1.** Ned Block, "Inverted Earth," *Philosophical Perspectives* 4 (1990): 53–79, esp. 53–54; restated and extended in "Mental Paint," in M. Hahn and B. Ramberg, eds., *Reflections and Replies: Essays on the Philosophy of Tyler Burge* (Cambridge, MA: MIT Press, 2003). The novelty over the classical inverted spectrum is the use of the representationalist's own externalism to drive content and phenomenal character apart, rather than relying on a bare intuition of undetectable inversion.

**2.** The reconstruction follows Block 1990, pp. 62–64, and the presentation in Block 2003. The two coordinated inversions — objects *and* color vocabulary — are essential; without the vocabulary inversion the locals would simply contradict the traveler, and the scenario would not cohere.

**3.** Block grants the arrival-stage alignment explicitly (1990, pp. 63–64); the argument needs the *temporal* stage, where externalism has had time to act, and does no work at arrival.

**4.** The externalism is Putnam's and Burge's: Hilary Putnam, "The Meaning of 'Meaning'," in *Mind, Language and Reality: Philosophical Papers, Vol. 2* (Cambridge: Cambridge University Press, 1975), 215–271, esp. 223–27. Block's dialectical point is that the representationalist cannot reject the externalism without abandoning the naturalistic account of content that motivates representationalism in the first place.

**5.** "Mental paint" is Block's term for intrinsic, non-representational phenomenal properties of experience; see Block 2003, esp. 165–66, and, for the access/phenomenal framing it leans on, "On

a Confusion about a Function of Consciousness," *Behavioral and Brain Sciences* 18 (1995): 227–247, esp. 227–30.

**6.** Michael Tye, *Consciousness, Color, and Content* (Cambridge, MA: MIT Press, 2000), sec. 6.2, "Biting the Bullet on the Inverted Earth Objection: A Reply to Block," 123–134. Tye introduces a possible hard-line response and grants the representational shift for that discussion. He then argues that memory-based judgments of unchanged character need not be authoritative. This is a defense available to the externalist, not an independently observed phenomenal shift; Section 6.3, 134–135, rejects the route over its indeterminate transition; Section 6.4, 136–140, gives Tye's own no-shift alternative, including accidental replicas. The book adopts the explored route while acknowledging this unresolved objection, not Tye's final conclusion.

**7.** The synch-shift reply — phenomenal character changing gradually in step with content, so that the two never dissociate and the change escapes the subject's notice — was pressed against Block by William Lycan (*Consciousness and Experience*, 1996) before Tye examined and then rejected it; Block had anticipated the move himself ("Inverted Earth," 68). Block characterizes and targets it in "Mental Paint" (2003), esp. 185–87: "The gradual shift of phenomenal character in sync with the gradual shift of representational content avoids any gap between them."

**8.** Tye, *Consciousness, Color, and Content* (2000), sec. 6.2; the phenomenal-memory datum ("when, on Inverted Earth, I remember the clear sky of my youth, my phenomenal memory image is of the blue earthly sky") is the standing pressure on the synch-shift reply. It records pinned *reference*, not preserved *character*: Tye holds in the same section that the image's being of the blue earthly sky is "compatible with supposing that the color it represents the sky as having is yellow," so that "phenomenally my memory of the sky matches my present visual experience" — the reconciliation Section 5.1 builds on. The Thatcher case supplies the pinning half (the image stays of the earthly Thatcher, whatever the subject now believes); the represents-as half re-tunes with perception's content. Block's complaint that externalism about memory amounts to an ad hoc error theory is in "Mental Paint" (2003); Chapter 6 states the answer in brief — the full exchange lives here. The object/property distinction explains the consistency of the proposed memory shift. Its phenomenal interpretation still depends on wide identity. The structural rival can agree that memory has wide accuracy conditions while holding its phenomenal character fixed by present organization; generic memory externalism does not exclude that position.

**9.** On color as objective surface-reflectance-type — the color realism the book defends — see Alex Byrne and David R. Hilbert, "Color Realism and Color Science," *Behavioral and Brain Sciences* 26 (2003): 3–21. Reflectance realism itself is neutral between the two representationalist routes: on either, experiential content fixes on objective reflectance-types; the routes differ only on whether the type a state tracks can re-anchor with the creature's history. The declined route reads the tracking relation as pinned by design and calibration; the book's consumer semantics lets it re-tune (Section 6, Objection 1).

**10.** Block, "Mental Paint" (2003), esp. 188–89, reporting the rival evolutionary-anchoring view (which he attributes to Dretske, Tye, and Burge) in order to resist it: "color representation is a product of evolution, hence if the sky on Inverted Earth continues to produce the same phenom-

enal experience in our traveler, it also is represented as blue." This is the route the book declines (Section 5.1) — though for the book's own reasons, not Block's: his resistance runs through perceptual adaptation; the book's runs through the possibility of learned consumer recruitment. Successful performance alone does not establish either new content or the accuracy of existing content.

**11.** Gilbert Harman, "The Intrinsic Quality of Experience," *Philosophical Perspectives* 4 (1990): 31–52, esp. 39; G. E. Moore, "The Refutation of Idealism," *Mind* 12 (1903): 433–453, esp. 446, for the diaphanousness observation; Michael Tye, "Representationalism and the Transparency of Experience," *Noûs* 36 (2002): 137–151, esp. secs. 2–3.

**12.** The "smaller room" is the point: relocating the residue from a sense-datum to a bare phenomenal property does not exempt it from the demand that a posited item do some work. Cf. the anti-reification methodology — mental terms name relations and processes, not inner objects. The anti-reification point must not be used to exclude a physicalist vehicle identity: such a theory can offer a distinct constitutive explanation while matching all predictions. A missed contrast would strengthen its case but is not a prerequisite for competition. Nor does this test count against the residue-free structural rival.

**13.** Ruth Garrett Millikan, *Language, Thought, and Other Biological Categories* (Cambridge, MA: MIT Press, 1984), supplies the consumer-side historical framework. Fred Dretske, *Explaining Behavior* (Cambridge, MA: MIT Press, 1988), chs. 4–5, develops learned indicator functions. His *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995), ch. 1, 12–18, expressly distinguishes systemic sensory representation from acquired conceptual representation. Learned recalibration of the latter does not automatically re-anchor the former. The book's proposal that consumer recruitment can change perceptual property content is its own synthesis, not a consequence already established by Dretske's distinction. Evidence of new payoff or new word use alone therefore cannot establish the perceptual shift; still less can it establish a change of phenomenal character.

**14.** Block, "Mental Paint" (2003), esp. 171–73; the awareness-without-attention cases and the limits of introspective absence are deployed against the transparency theorist's appeal.

**15.** Block 2003, esp. 165–66 and 184–89. The inverted-spectrum and Inverted-Earth cases are presented as cumulatively establishing that intentional content underdetermines phenomenal character.

**16.** Tye, *Consciousness, Color, and Content* (2000), sec. 6.2, 123–134, develops the memory-based defense of synch-shift. Sections 6.3–6.4, 134–140, then reject that route and propose no sensory shift, including for swamp replicas. The book adopts the route explored in 6.2 but does not attribute its adoption to Tye; his transition objection remains a debt, stated in Section 5.1. Memory externalism defends the route against a simple appeal to remembered sameness without independently establishing phenomenal change.

**17.** William G. Lycan, *Consciousness and Experience* (Cambridge, MA: MIT Press, 1996). On the tracking-representationalist family (Lycan, Dretske, Tye) and its shared reliance on causal-

historical content-fixing, see Lycan, "Representational Theories of Consciousness," *Stanford Encyclopedia of Philosophy*.

**18.** David J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press, 1996), esp. ch. 3, sec. 2, on the coherence of inverted spectra.

**19.** Tim Crane, "The Origins of Qualia," in Tim Crane and Sarah Patterson, eds., *History of the Mind-Body Problem* (London: Routledge, 2000), 169–194. Crane treats Inverted Earth as a focal case in the qualia/representationalism dispute.

**20.** Tye announced the panpsychist turn in *Vagueness and the Evolution of Consciousness: Through the Looking Glass* (Oxford: Oxford University Press, 2021), distinguishing sharp fundamental consciousness* from the determinate, representationally fixed forms of consciousness that may inherit vagueness. He restated and revised the case for a general audience in "How I Learned to Stop Worrying and Love Panpsychism," *Journal of Consciousness Studies* 31, no. 9–10 (2024): 10–28: the sharpness argument appears at 13–20, and the consciousness*/content plus global-workspace transfer proposal at 20–24. The two presentations belong to one evolving view but should not be collapsed into the simple claim that all ordinary consciousness is sharp.

# Appendix B — Scholarly Excurses

*Deep-dives the general-reader spine deliberately omits. Each chapter states its verdict in brief; the extended treatment lives here.*

### B.1 Inner Speech and the Vygotskian Story

*(Deep-dive from Chapter 17, The Assembled Self.)*

Right now, as you read this, the words on the page sound, silently, in something you would call your own voice. Whatever else this thing is, it feels like the deepest interior we have: a private room with the door shut, the last place the world does not get in. This excursus is about why that picture runs upside down.

The bad picture says public speech is the noisy externalization of an already-private inner monologue: the thoughts come first, in the head; the words come later, when the thoughts need company. The picture comes naturally, and it has done more than almost any other to lock the Cartesian theater into modern philosophy of mind. Once you accept that the inner voice is *the thing*, you have already conceded an interior space with its own contents, accessible only to its owner.

The alternative runs the other direction. Public speech came first, historically in the species and developmentally in each child, and the inner voice consists of that public activity turned inward. The developmental half of that claim is Vygotsky's story: the child talks with others first, then to herself aloud, then silently, speech migrating inward by stages rather than thought leaking out. What you experience as thinking-in-words is silent speaking in a transformed register: the same lexicon, the same public meanings, the same shared language — transformed, not replayed at low volume. Vygotsky's own finding was transformation-in-internalization: inner speech abbreviates, drops its subjects, condenses sense, runs in a predicative shorthand no listener could parse. (The mechanism beneath the phenomenology — attenuated motor commands, auditory-forward imagery, some blend — stays contested in the psychology, and nothing philosophical here turns on it.) No separate inner lexicon exists. One lexicon (public, shared, social) does duty in two modes, the inward mode compressed.

Several puzzles relax their grip once you see this. Take meaning. If the inner voice were a soliloquy in a private language, its words would have to be fixed inside the head, by the speaker's own lights, with no public check — exactly what Wittgenstein attacked in the case of someone trying to name a private sensation, where no public criterion of correct use can settle whether the next use follows the rule or breaks it.[1] Meaning never gets fixed by what goes on inside a single skull; it gets fixed in social practices where words have uses people can correct, share, and inherit. The inner voice borrows those meanings from outside. It does not generate them. Ask the obvious question: where else would it get them? You learned every word you have from speakers around you, in contexts where their uses could be corrected.

Here we owe Sellars and the tradition after him. His Jonesean myth models inner episodes — thoughts — on overt verbal utterances rather than treating them as directly observed.[2] Crane and Farkas gloss the picture: thoughts are "inner episodes, called 'thoughts,' which are conceived on the model of overt verbal utterances, but happen silently in the head of other subjects."[3] The order

of explanation runs from public to private. We learn what thinking is by learning to talk; *then* we learn to call the silent version "thinking."

Two consequences follow. The first concerns AI. If the inner voice is internalized public speech, then producing fluent inner-voice-like sentences does not, by itself, establish an inner life behind them. A language model has the surface of inner speech and none of the social, embodied developmental history through which a human speaker learned to use those public meanings; what training can select inside the model's mechanisms is a further and separate question (note 4, and B.15 below).[4] Causally internalized public meaning does not exhaust that question: where the training history Mallory describes obtains, content about actual word-types and their distributions is original at its own grain, as B.15 explains, without settling reference or understanding. The temptation to see understanding behind the outputs comes precisely from the inverted picture: we assume the words must come from an interior, because that is where our own words seem to come from. They do not. The second consequence concerns us. The most private-feeling activity we have is, at root, a social inheritance — the point Chapter 17 presses at length: every word in the monologue arrived from outside, on loan from a community.

An obvious objection: surely there is more to thinking than silent speech. Mathematicians solve problems without verbalizing them; animals without language clearly think. Take it seriously, because it is mostly right — and the claim survives the objection once narrowed. Not every act of cognition consists in inner monologue; pre-linguistic infants think, non-human animals think, and expert performance outruns any inner narrator. But the *experience* of thinking-in-words, the voice that feels like the inmost private chamber, is best understood as internalized public speech.[5] Non-verbal cognition is real; it is simply not what feels private. The room you thought was private always had the door open. You just never noticed the draft.

### B.2 Information Talk Is a Metaphor

*(Deep-dive from Chapter 24, The Conscious Robot.)*

Open any popular book on the brain and somewhere in the first chapter you read that the brain *processes information* — said with the same equanimity as *the kidneys filter blood*. Information goes in through the senses, the brain does things to it, behavior comes out. The picture seems so plainly right that the only response is to nod and turn the page. Look at the sentence again.

What exactly enters the brain? Not the photons; those stop at the retina. Not the air-pressure waves; those die in the cochlea. What travels into the skull does so as ion flux across neural membranes — sodium in, potassium out, at rates no ledger could follow. The neurons carry no envelopes marked *information* any more than a river carries envelopes marked *fluid dynamics*. Somewhere between *the neurons fire* and *the brain processes information*, something has been added, and it pays to ask what.

The deflationary point begins where Chapter 22 left off, but it must begin on the right side of that chapter's distinction. Unconstrained computational interpretation is observer-relative; disciplined implementation can be an observer-independent fact about a mechanism's causal organization. The same care applies to *information*. Shannon information measures real statistical dependence. Semantic information adds a further claim about what a state is about. A neuron

fires; another fires in response; the causal and statistical relations belong to the tissue. Calling the second state information *about an edge* or *about food* requires more: a history, selected function, and producer-consumer use that privilege one correctness condition over its rivals. Searle puts the warning at its most disarming: "'Information' does not name a real physical feature of the real world in the way that neuron firings, and for that matter consciousness, are real physical features of the world. Except for the information that is already in the mind of some conscious agent, information is relative to an observer. […] Information is anything that we can count or use as information."[6] His target remains the slide from a useful reading to intrinsic semantics. The book grants more causal structure than his warning acknowledges: selection and training can produce observer-independent, content-bearing mechanisms. What it refuses is the inference from causal traffic, computation, or measured bits to semantic content without a privileged explanatory mapping.

Notice what has *not* been said. Nothing against describing the visual system as extracting edges or the hippocampus as encoding locations, and nothing against those descriptions being literally correct when selected functions fix the content. The point cuts between a relation and a substance. *Information* does not name stuff the traffic contains. The grammatical mistake runs ahead of the metaphysical one: *processes information* suggests handling, handling suggests cargo, and cargo lives somewhere — little packets of meaning piling up in the visual cortex like luggage in a baggage claim, waiting for someone to collect them. Cue the inner theater.

Now the objection that wants it all back: Shannon. Surely there we have information in a precise, mind-independent sense — bits, channels, probability distributions, all perfectly real. The objection deserves a careful answer, because most of it is correct. Shannon information exists and gets measured; Dretske built an epistemology on it.[7] But two points. First, Shannon information is not semantic information: it measures statistical dependence between source and channel, the same dependence that holds between a room's temperature and the height of mercury in a thermometer. (Dretske's *natural information* — factive correlation, what a state's obtaining guarantees about its source — sits between the two: stronger than Shannon's averaged quantities, still short of a correctness condition a state could fail to meet.) Statistical covariance is real; content needs something further to select one condition as what the state is correct about. Second, the slide from computability or simulability to a substantive theory of mind repeats the Church–Turing fallacy Chapter 22 diagnoses.[8] A brain may genuinely implement computations; that empirical result still does not show that computing alone constitutes thinking. Shannon attributions come nearly free. Semantic ones cost the right selected function, causal embedding, consumer role, and environmental history. Conflating the two gives the illusion of having explained meaning by measuring bits.

The bad picture is not that we talk about information in the brain; we can hardly do without the talk. The bad picture is that the talk names a stuff. It does not. Shannon information names a statistical relation; semantic information names a representational relation whose correctness conditions are fixed by function, history, and producer-consumer organization. The organ stays biophysical through and through. No packets required.

### B.3 The Conceivability Trap

*(Deep-dive from Chapter 9, Why Dualism Keeps Winning Arguments It Should Lose; the strong-necessity exchange deferred from Chapter 11 completes it.)*

Imagine a being physically identical to you in every respect. Same neurons, same synapses, same electrochemical cascades firing in the same sequence. It winces when burned, says *ouch* with conviction, and passes every behavioral test. The only stipulated difference: nothing it is like to be this creature. No red when it sees the tomato, no warmth when it holds the coffee. The machinery runs in darkness.

This philosophical zombie gives David Chalmers his sharpest modal argument against physicalism.[9] The popular summary says: if we can imagine the zombie, it must be possible. Chalmers's argument says something narrower and much stronger. Casual imagery supplies no premise. The relevant achievement is *ideal primary positive conceivability*, and the possibility it reaches is *primary possibility*.

#### The argument in its exact form

Let **P** state the complete physical truth about a world and **Q** a phenomenal truth, such as *someone is conscious*. Levine's explanatory gap supplies the setting: P can fail to make Q intelligible even if P and Q describe one fact.[10] Chalmers adds this modal argument:

1. **P & not-Q is ideally primarily positively conceivable.** One can coherently imagine a sufficiently detailed scenario that verifies P & not-Q when considered as actual, and the imagining survives ideal rational reflection.
2. **Ideal primary positive conceivability entails primary possibility.** If a statement passes that test, some centered metaphysically possible world, considered as actual, verifies it.
3. **The phenomenal case has no water-style primary/secondary slack.** Core phenomenal concepts do not fix their reference through a contingent description such as *the watery stuff around here*; their primary and secondary intensions coincide.
4. **Therefore P & not-Q is secondarily, metaphysically possible, and physicalism is false.** A metaphysically possible physical duplicate lacks the phenomenal truth, so the physical facts do not necessitate all the facts.[11]

Each qualifier in the first premise blocks an easy physicalist reply.

*Ideal* removes errors that better a priori reasoning would defeat. A mathematical impossibility may look coherent to us because we have not found the proof; it does not remain conceivable under the idealization. *Positive* requires coherent modal imagination of a situation that verifies the statement. Merely supposing a sentence, failing to detect a contradiction, or picturing experts announcing it does not suffice. Positive conceivability need not involve a visual image. The zombie and its conscious twin look the same from outside, yet a thinker can still modally imagine a world-configuration and take it to verify that one lacks experience. *Primary* asks how a world might turn out when considered as actual. It differs from asking, after the actual referent has been fixed, what could happen to that very referent in a counterfactual world.[12]

That last distinction answers the standard water objection. "Water is not $H_2O$" has a primary reading on which XYZ fills the lakes and rivers in a world considered as actual. Such a world is possible and verifies *the watery stuff is XYZ*. It does not show that actual water, whose reference science fixed as $H_2O$, could have been XYZ. Primary conceivability reaches a genuine possible world; secondary necessity keeps actual water $H_2O$ in every counterfactual world. The Kripkean case therefore supports Chalmers's two-dimensional bridge rather than refuting it.

The phenomenal case, Chalmers argues, lacks the same escape. A core phenomenal concept does not pick out its referent through a contingent description analogous to *the watery stuff around here*. What verifies *pain* or *red experience* presents the phenomenal property itself. Primary and secondary intensions therefore coincide. If an ideal primary scenario verifies P & not-Q, no later redescription turns it into a merely watery duplicate of the target. The scenario is a zombie world, and physicalism loses.[13]

**Where type-B physicalism refuses the bridge**

The book grants premise 1. Grant it ideally, primarily, positively, and permanently. No future scan closes the conceptual distance. The acquaintance-recognition account predicts that permanence: a concept acquired through undergoing a state need not acquire an a priori connection to a concept built from public description. Ideal reflection improves reasoning; it does not rewrite the acquisition history of the concepts or turn one epistemic route into the other.

The refusal lands on premise 2. Ideal primary positive conceivability supplies serious modal evidence, not an entailment of primary possibility. Modal judgment should remain answerable to independently supported identities. If Chapter 6′s identity holds, the acquaintance-grounded and descriptive concepts reach one physical world-presenting state even though no ideal a priori route joins them. P & not-Q can then remain ideally positively conceivable under the separated concepts while failing to describe a metaphysically possible world. The scenario records conceptual isolation rather than ontological separation.

This answer accepts a real price: strong necessity. It offers no neutral counterexample to Chalmers's restricted principle. Its positive ground lies instead in Chapter 6′s independent identity case[14] and the comparative costs counted in Chapters 9 and 10. The remaining standoff concerns which source of evidence governs when ideal conceivability and an abductively supported identity conflict. This book chooses the identity and says so plainly. Chalmers's argument has not been defeated; its decisive premise has been refused for a reason supplied elsewhere in the book.

**What the intuition tracks**

The zombie intuition still calls for explanation. Chalmers's meta-problem asks why creatures like us judge consciousness so puzzling, and why we produce the reports and arguments we do.[15] The acquaintance-recognition account answers part of that question. A physical description can predict the gap-talk without conferring the first-person route whose absence the talk records. The inner-theater picture adds a second influence: once experience gets pictured as a private item over and above the representational work, subtracting the item while leaving the machinery intact feels natural. That picture helps explain why the modal bridge attracts us. It does not show that the conceiving fails.

One intuition outlasts every diagnosis: pain hurts, and a description of nociception does not. Correct. A description describes; it does not put the reader in the foot, the wince, or the moment. A weather report does not get you wet. The missing transmission marks the difference between describing a state and undergoing it. Converting that difference into a second property requires the very modal bridge under dispute.

**Two responses worth naming**

The ability hypothesis associated with Lewis and Nemirow answers Mary's knowledge argument, not the zombie argument.[16] It may complement the present account, but no thesis about Mary's new abilities settles the conceivability-to-possibility inference.

Loar, Papineau, Balog, and others develop the broader phenomenal-concept strategy: phenomenal and physical concepts can remain cognitively isolated while referring to one physical property.[17] The house account adopts a narrower mechanism. Acquaintance with an occurring world-presenting state grounds recognition; description alone does not confer that route. Chapter 11 and B.8 state the limit without camouflage: this mechanism explains the form and permanence of the gap if identity holds. It neither establishes identity nor independently defeats Chalmers's modal principle.

The ledger therefore closes narrowly. The zombie stays ideally primarily positively conceivable. Chalmers reads that rational achievement as primary possibility. This book reads it as defeasible evidence outweighed by the independent identity case. The gap remains epistemic; the identity keeps the ontology physical.

### B.4 The Church–Turing and Church–Turing–Deutsch Theses

*(Deep-dive from Chapter 22, Simulation and the Observer [Sections 22.5 and 22.6].)* Two celebrated results stand behind much of computationalism's public confidence, and both get cited as though they settled questions they do not address. They are siblings (one bounds the mathematics, the other concedes the physics), and the rhetorical pattern that exploits them depends on never examining either too closely. This excursus states each accurately, walks the inference each gets drafted into, and itemizes what the computationalist still owes once the borrowing stops.

**The Church–Turing thesis (1936).** Church and Turing, publishing independently, gave precise definitions of what it is for a function to be effectively computable — workable out, in principle, by following a finite list of mechanical rules — and showed their definitions equivalent. The thesis asserts that this formal notion captures the intuitive one: any function a clerk with paper and pencil could in principle grind out by rule is computable by a Turing machine. That is the whole of it. It speaks of functions and formalisms. It does not speak of brains, or minds, or physical systems at all.[18]

The public-facing argument nonetheless runs: anything computable can be computed by a Turing machine; the brain is a physical system; therefore the brain is a computer, and a good enough program will eventually be a mind. Watch the inference try to cross the gap. *The brain computes things* — but if that means it manipulates discrete symbolic states by rule, neuroscience has found no such thing; that is the commitment in dispute. And if it means only that some computable function can *describe* the brain's behavior, it is harmless and trivial, true in equal measure of a

hurricane, a stomach, and the solar system. *Therefore the brain is a Turing machine* — but the thesis says nothing about which physical things are Turing machines; it fixes what functions such machines compute, not which lumps of matter qualify as one. *Therefore running the program would be a mind* — the most ambitious leap of all, from *the brain can be modeled by a computation* to *running the model would constitute the thing modeled*. The thesis is silent on every step that matters. Jack Copeland named the pattern the *Church–Turing fallacy*; Gualtiero Piccinini supplied its anatomy, and showed several historically influential versions of it — Newell, Pylyshyn, and others — to be unsound.[19]

To get from the theorem to computationalism, the computationalist needs three further things the thesis does not give him. He needs a substantive account of when a physical system genuinely computes, one that tells the brain apart from the rock. He needs an empirical case that the brain, on that account, computes. And he needs a philosophical argument that mental life is *constituted* by such computation rather than merely *described* by it.[20] Each is a separate undertaking. None follows from the theorem. Stripping the thesis of the role it was press-ganged into does not refute computationalism; it returns the burden of proof to where it always sat — on the empirical and conceptual case for treating cognition as computation, to be fought on its own ground.

**The Church–Turing–Deutsch principle (1985).** The thesis's physical sibling, due to David Deutsch, is a proposed physical principle rather than a theorem: every finitely realizable physical system can be *perfectly simulated* by a universal model computing machine operating by finite means.[21] Deutsch's own accounting is stricter than the slogan that circulates: classical physics and the classical Turing machine do not satisfy the principle in this strong form — it took the universal *quantum* computer to make it physically credible — and the looser "to arbitrary accuracy" formulation is a later weakening. Granting the principle here therefore hands the computationalist the strongest simulability premise on offer. The brain is a finitely realizable physical system; therefore the brain is simulable. The physicalist should grant this, or very nearly — denying it means betting on a non-computable physics, a bet a few serious people have placed (Penrose most famously, on quantum-gravitational grounds) but not a debt the computationalist's opponent has any need to take on.

Granting it costs nothing, because the principle delivers *simulability* (the existence, in principle, of a formal model that reproduces the system's behavior), and simulability lives squarely on the representation side of the representation/instantiation line. "The brain is computable" says a description exists; it does not say that running the description anywhere is the thing described. The gap between *X is computable* and *the computation of X is X* is the gap Chapter 22′s hurricane opens, restated in the vocabulary of a theorem — and a theorem about the existence of a model cannot, on pain of changing the subject, close a question about whether the model is the thing.

Set side by side, the two results expose the oscillation that does the real argumentative work. Where the thesis bounds the formalism, the principle concedes the modeling: one says *these are the computable functions*, the other says *the brain is among the simulable systems*. Neither carries one inch from model to mind. The computationalist's habit is to claim the strong thing when convenient (being a mind just *is* running the program) and to retreat, when pressed, to the weak one: the brain is computable, the brain is simulable. And he hopes no one notices that

the conclusion always required the strong version while only the weak was ever earned. The mathematics and the physics both survive scrutiny intact. The metaphysics they were made to carry has to walk on its own feet.

### B.5 Color and the Reflectance Account

*(Deep-dive from Chapter 6, The Identity Claim.)*

Chapter 6 defends two claims that need not stand or fall together. Phenomenal character consists in a conscious state's world-presenting profile; successful perception reaches the tomato itself rather than an inner proxy. Neither claim alone settles whether tomatoes really have the colors they appear to have. You can see a thing while getting a property of it wrong. The book adds a third claim, defended here: ordinary surfaces really are colored, and their colors consist in types of spectral reflectance. That account gives the visual content a physical property to answer to. It earns its place by explaining color perception, not by following from the identity claim.

That sounds like the blandest common sense, and the standing surprise of the subject is how many careful people have felt driven to deny it. Pull the wavelengths apart and the worry sets in. Physics, examining the tomato, finds a surface that reflects light of certain wavelengths and absorbs the rest; it finds no further property, no *redness*, riding on top of the reflecting. Where, then, is the red? Three answers tempt, and the book turns down all three.

The boldest answer says: nowhere. Nothing has the colors ordinary vision attributes to surfaces. C. L. Hardin brings the empirical study of color to this error-theoretic challenge; Paul Boghossian and David Velleman argue that experience presents colors as intrinsic qualities which physical surfaces lack.[22] Such a view need not put a colored picture in the head. It can say that we see real tomatoes and systematically misrepresent their color, just as we can see a real stick and misrepresent its shape. Global error about one property does not abolish perceptual contact.

The dispute concerns what vision attributes. If color experience promises a simple quality over and above every physical property of the surface, the error theorist has a powerful case. The reflectance realist denies that promise. A surface need have no extra redness riding on top of its physical disposition to modify light. The question is whether a reflectance-type can account for the property vision presents, including its stability across changing conditions. We must answer that question rather than make error theory incoherent by definition.

A gentler answer keeps color real but moves it into a relation to us: red just *is* the disposition to look red to normal perceivers in normal light — Locke's secondary quality, refined. This saves the appearances and pays a quiet price: it defines the witness by the testimony. The worldly property the experience was supposed to reach turns out to be specified in terms of that very experience, so the redness *out there* smuggles the perceiver back inside the object. For a thesis trying to anchor felt character in something genuinely world-side, that circle runs the wrong way.

That familiar dispositionalism does not exhaust the relational alternatives. A thicker ecological view treats red as the surface's stable way of interacting with illumination, surrounding colors, movement, and a class of perceivers across the relevant transformations. The property belongs to the surface-in-an-environment rather than to an inner response. Dennett's claim that colors and

color-vision systems were "made for each other," and Noë's account of colors as patterns in how surfaces vary in appearance under changing conditions, both give this rival objective accuracy conditions without private paint. Transparency sits comfortably with it: the relation appears as a way the surface is, not as a colored item in the head. The crude circularity charge therefore cannot decide the issue.[23]

A third answer honors the phenomenology hardest. Red, it says, is just what it looks like — a simple, irreducible quality the surface presents, with no room for it in the physics: not a reflectance, not a disposition, but a *sui generis* feature standing there in its own right. John Campbell defends a version under the disarming name of "a simple view of colour."[24] The book rejects this additional irreducibility in favor of a physical account of surface color. The question is whether primitive color explains something the physical account cannot. Calling it a second mental realm would prejudge that dispute: Campbell puts the color in the world, not in a private mind.

The reflectance account combines three commitments whose conjunction these alternatives dispute. Color stays real, against the error theory; it belongs to the surface and not to us, against dispositionalism; and it sits squarely inside the physical world, against primitivism. Red just *is* a type of surface spectral reflectance — the tomato's standing physical disposition to throw back a certain profile of wavelengths and soak up the rest. Alex Byrne and David Hilbert, who argue that color science vindicates color rather than debunking it, work the view out in detail; Tye combines reflectance physicalism with his representationalism, and the book defends that combination on its own terms. One wrinkle disciplines the view. Many physically different surfaces look the same red (*metamers*, reflectance profiles that part ways on the spectrophotometer and match to the eye), so a color cannot be one reflectance. It is a *type*: a family of reflectances our visual system groups together as a single color. Here the dispositionalist will cry foul. Has the perceiver not just been smuggled back into the object — the very move the book objected to when it turned dispositionalism down? No, and the difference carries the whole weight. Our responses *select* which physical properties to file under "red"; they do not *constitute* the properties filed. Each reflectance in the family is a perceiver-independent physical fact about how a surface treats light, there whether or not anyone looks. Human vision contributes a biologically constrained partition, shaped by coevolution and sensorimotor organization, that groups those facts into colors. Dispositionalism makes the color *be* a relation to us. Reflectance realism makes the color a physical property while allowing an evolved visual system to constrain which physical families count as colors. That anthropocentric individuation is explanatorily important rather than arbitrary bookkeeping. We carve physical kinds to human interest all the time ("weed," "continent," "tropical") without the carved kinds ceasing to be physical, and the result is a naturalistic color realism that can call the tomato red without finding anything in it over and above its physical surface.

The ecological rival nevertheless identifies a real feature of color vision. A visual system does not recover a spectrophotometer's curve; it tracks a stable surface through changes in illumination, viewpoint, comparison, and action. Reflectance realism should treat that counterfactual profile as the route by which perception groups and tracks reflectance-types, not pretend the route contributes nothing. A complete color experience presents two coupled features: an invariant object-color and its current manifestation under this illumination, surround, viewpoint, and visual

mode. The gray patch may look pinkish against one field and dark gray against another while retaining one surface color; the tomato may change appearance between garden and fluorescent light while vision continues to track it as one red. Those contextual differences need not count as errors. Color constancy gives the preference its point: reflectance supplies the distal stability that remains while proximal light and current manifestation vary. The ecological realist takes the transformation profile as partly constitutive of color; the book treats it as the organization of access and appearance through which a stable color gets presented. Constancy supports that division without proving it. Where two reflectance families sustain the same complete ecological profile for a kind of perceiver, the view should concede the same apparent shade and may group them under one anthropocentric color type. It need not invent a phenomenal difference at a finer microphysical grain. Nor must the color property contain its own appearance. Strong representationalism places phenomenality in the conscious state's presenting of the property, under a visual mode, rather than requiring redness itself to look red when no one sees it.[25]

Metamerism is only half of the empirical challenge; the other half runs in the opposite direction. Instead of different reflectances producing matching appearances, the same surface can appear different to different observers. Normal observers disagree about which chips look *unique green*: green untinged by blue or yellow. Hardin presses such variation against perceiver-independent color. Coarse reflectance-types offer part of a reply. Observers may correctly attribute the same broad type while differing in finer hue presentation. But calling those differences errors requires independently specified fine-grained accuracy conditions. It also requires distinguishing error from context-sensitive appearance. Coarse typing alone settles neither whose unique-hue judgments are correct nor how often observers misperceive a shade. Reflectance realism must meet that finer-grained challenge; the fact that the surface has a stable reflectance does not answer it.

The reflectance account must also face the inverted spectrum, one of the oldest objections in the color literature. Imagine your color experience is systematically inverted relative to mine — what I experience seeing red tomatoes, you experience seeing grass, and the reverse. But we were both trained from childhood to call tomatoes "red" and grass "green," so our verbal behavior is identical; we both say "that tomato is red" and mean it. Is there a difference between us? Intuitively many say yes: your experience of tomatoes feels, from the inside, the way mine of grass does. You have inverted qualia. If the scenario is coherent, then phenomenal character cannot consist entirely in representational content, for our experiences represent the same properties while feeling different.

This deserves a real response. The case asks whether phenomenal character could come apart from content. Its apparent conceivability supplies a challenge, but does not by itself establish the metaphysical possibility the objection requires. That distinction gives the identity theorist work to do, not permission to dismiss the example.

A stronger response goes deeper. The full content of a color experience is not just which surface property gets attributed to which object; it includes the color's *relational* properties — its location in color space, its similarities and differences to other colors, the way it figures in categorization. Red and green are not arbitrary labels; they differ in how they relate to each other and to the rest of the color solid. If my experience represents red in that full relational sense (as the color that

contrasts with green, resembles orange, and not blue) then your genuinely inverted experience would have to represent those relations differently too, and then it might not represent the same properties at all. Inversion may not be coherent for a system as integrated as human color vision. That is not a refutation; it is a reason to doubt how much weight the thought experiment bears. It shows we can *describe* a scenario in which character comes apart from functional role; it does not show the scenario is a genuine metaphysical possibility rather than a description that unravels on examination. (The inversion case is Sydney Shoemaker's, "Functionalism and Qualia," *Philosophical Studies* 27 [1975]: 291–315; the representationalist response is developed in Michael Tye, *Ten Problems of Consciousness*, ch. 7, and Alex Byrne, "Inverted Qualia," *Stanford Encyclopedia of Philosophy* (2004; substantive revision 2025), with the relational-structure move stated most clearly in Gilbert Harman, "Explaining Objective Color in Terms of Subjective Reactions," *Philosophical Issues* 7 [1996]: 1–17.)

On this combined account, the tomato supplies a worldly anchor for color vision. Its surface carries the reflectance-type a successful perception presents. The felt redness consists in that conscious presentation, not in an additional red item behind the eyes. An illusory presentation can have the same character while attributing a property its object lacks; accuracy determines whether the color attribution succeeds, not whether the subject has an experience. Reflectance realism and phenomenal identity thus do different work. One identifies the worldly color; the other identifies the character of its presentation. The book defends both without treating either as proof of the other.

### B.6 The Two Objections to the Mark

*Chapter 1 states the mark and leaves its objections here. This excursus carries the full dialectic: the two apparent counterexamples that have shadowed Brentano's thesis for a century and a half, the inner-theater temptation they invite, and the harder case they leave open.*

#### The Mood That Looks Like a Counterexample

Sit with a diffuse, free-floating anxiety, not a worry bolted to a deadline but the kind that settles over a gray Sunday afternoon and will not say what it wants. Ask what it is *about*, and the answer comes back *nothing in particular*. The dread just sits there, apparently aimed nowhere. Searle concedes the point in the first paragraph of his *Intentionality*: beliefs and desires are directed, but "there are forms of nervousness, elation, and undirected anxiety that are not Intentional."[26] He does offer a reporting-constraints test in the sentence that follows. Crane's narrower complaint still lands: Searle treats the existence of non-intentional mental states as too obvious to require real argument.

The mood lacks a *particular* object. Whether it lacks content remains the disputed point. Some intentionalists read anxiety as directed at the world taken whole, where threats loom and the day narrows; others take it to present the felt condition of one's own body. Angela Mendelovici declines the object hunt. Undirected moods, she argues, really do lack an object, and insisting that they aim at some unusual one gets their phenomenology wrong. They carry content regardless: an unbound affective property, a badness felt without fastening to anything.[27] Her case costs

the mark its tidiest object-directed formulation while preserving the content-having claim the formulation served.

Anxiety and elation can turn us toward the world in different ways. That gives an intentionalist something to explain, but the difference could arise from contentless affect changing other contents. Chapter 7 favors a body-and-field account without claiming to have ruled out that rival. The mood's apparent lack of an object therefore fails to refute the identity thesis; it does not confirm the thesis either.

**The Monster That Isn't There**

The second objection comes from the other direction. A child lies awake certain that a monster crouches under the bed, large and somewhat damp. The fear is real. The monster is not. Yet "nothing" gives the wrong answer to what the child fears: the state has a sharp and specific content, *a monster under the bed*.

Brentano revived a Scholastic phrase for the object's status, *intentional inexistence*. The expression has misled generations of readers. Its *in-* marks the object's existing-*in* the mental act, its immanence, not its nonexistence. The monster teaches the plainer point with which that doctrine is often confused: a thought's content needs no existing thing outside to answer to it.[28] "Fear of a monster" carries no entailment that some monster gets feared. Intentional description does not behave like an ordinary two-place relation whose two ends must both exist.

Meinong gave everything thinkable a standing short of existence: golden mountains, round squares, objects beyond being yet available to thought. Russell's theory of descriptions sought the same explanatory result without the population boom. The fear has propositional content, and that content stays determinate whether or not the world supplies a monster.[29] Nor does the case show that content floats free of the world. What makes this fear a fear of monsters, and what allows it to misrepresent, comes from a history of dealing with an actual world.

The more tempting repair does greater damage: relocate the missing monster inside, as a vivid image on a private stage. Once that move fixes the absent monster, every immediate object becomes an inner image, including the dog the child really sees. The world recedes behind a screen of pictures. Refuse the move and the objection leaves the mark where it found it. Aboutness needs no little item, inner or outer, to fasten onto. Aboutness is the fastening.

**What the Two Cases Leave Open**

The mood has no particular object and may still carry content; the monster has no existing object and the fear keeps its definite content. Neither case forces us off the mark. Neither closes the ledger.

Bodily pain presses harder. The toothache, the nausea, and the ache that seems to point nowhere answer to neither an objectless mood nor a missing monster. A pain may present a bodily region as damaged, or carry content that commands rather than describes. Which account the case requires belongs to Chapter 7; that phenomenal hurting consists in content of some such kind is the book's answer, and Chapter 7 must earn it. Chapter 1 needs only the narrower result: its two most familiar counterexamples do not yet defeat the mark.

### B.7 Emergence: The Wrong Question, Extended

*(Deep-dive from Chapter 11, The Gap That Isn't There.)* Section 11.5 separates the emergence of conscious organization from the supposed production of a further phenomenal ingredient. The weak/strong distinction, the British emergentist tradition, and the arguments of Crane and Tye help locate that boundary.

A traffic jam forms although no commuter plans one. A flock wheels although no bird directs the whole. Calling these patterns *emergent* need add nothing to the cars or birds and their interactions. Likewise, asking how neurons develop and sustain conscious, world-directed organization poses a legitimate scientific question. An identity claim cannot answer it by changing the vocabulary.

The disputed question comes later. Once that organization exists, must it produce a further feel? The book answers no: phenomenal character consists in the conscious state's complete world-presenting profile, one and the same thing under two descriptions. The organization still needs explaining. What the identity denies is an additional production step from that profile to its character. Two senses of emergence must therefore stay apart.[30]

*Weak emergence* concerns high-level patterns fixed by lower-level organization without new fundamental ingredients. Their novelty or computational complexity can make them surprising and difficult to predict. Explanations at the higher level may remain indispensable. A physicalist can accept all of this, including a weakly emergent organization of conscious capacities. The point does not require claiming that a bare physical description yields phenomenal knowledge *a priori*: the book's type-B position already denies that inference.

*Strong emergence*, in the sense opposed here, requires an irreducible addition to the physical account. The British emergentists developed influential versions of that proposal.[31] Some versions add fundamental downward causal powers; others posit basic psychophysical laws without such powers. Chalmers explicitly separates strongly emergent qualities from strong downward causation. Either addition for consciousness would contradict the book's physicalism, but they are different additions. Nor does failure of conceptual deduction establish either one. That would turn an epistemic limit into an ontological verdict before the argument had begun.[32]

Tim Crane complicates the boundary further. His comparison of emergentism and non-reductive physicalism questions how much metaphysical distance their different labels establish.[33] The useful questions concern what a theory adds: a new law, an irreducible property, an independent causal power, or only another explanatory description. Calling a view emergentist settles none of them. A residue-free grounding view also remains possible without an extra phenomenal ingredient; Chapter 6 addresses its proposed dependence relation separately.

Tye's trajectory brings out the attraction of the stronger demand. In *Ten Problems of Consciousness*, he objects to phenomenal consciousness arising in evolution with no explanation of its emergence.[34] His later panpsychism places consciousness among the fundamentals instead. The book declines both a brute phenomenal addition and fundamental consciousness. But those refusals do not amount to an explanation by themselves. They depend on the independently defended identity: if the feel consists in organized presentation, its occurrence requires the organization, not a further law which adds the feel to it.

There remains plenty for science to discover about how matter becomes organized into conscious systems. The identity claim denies no such development and explains none of its mechanisms by fiat. Its narrower payoff concerns the completed profile and its character. If the identity holds, asking how the first produces the second counts one fact twice. If it fails, the word *emergence* will not repair it.

### B.8 The Phenomenal-Concept Dilemma

*(Deep-dive from Chapter 11, The Gap That Isn't There.)* Section 11.7 gives the body-level verdict. This excursus states Chalmers's C/E dilemma in its topic-neutral form, identifies the horn the book takes, and adds Goff's separate opacity/translucency challenge.

#### Chalmers's master dilemma

Let **P** state the complete physical truth. Let **C** state, in topic-neutral psychological terms, the feature invoked to explain our epistemic situation regarding consciousness. Let **E** state that epistemic situation, also topic-neutrally: the truth-values, justification, and cognitive significance of the beliefs corresponding to our phenomenal beliefs. C cannot simply mean *phenomenal acquaintance*, and E cannot simply mean *knowledge of phenomenal facts*, because either formulation would build the contested phenomenology into the premise.[35]

A successful phenomenal-concept strategy needs three claims. Humans have C. C explains E. P physically explains C. Chalmers argues that the last two cannot hold together.

1. If **P & not-C** is conceivable, P does not transparently explain C. The explanatory gap has migrated from consciousness to the special feature of phenomenal concepts.
2. If **P & not-C** is not conceivable, any conceivable physical duplicate satisfies C. The zombie has the topic-neutral recognitional architecture, quasi-phenomenal concepts, and corresponding judgments.
3. Yet **P & not-E** remains conceivable. Zombie Mary's new dispositions and indexical judgment need not share the truth, justification, or cognitive significance of Mary's knowledge of what seeing red is like.
4. Therefore C either lacks transparent physical explanation or fails to explain E.

This differs from a dilemma between deducible *reports* and non-conferred *acquaintance*. Reports matter because they form part of the epistemic situation, but E includes epistemic status. Chalmers's second horn asks why the same C occurs in Mary and Zombie Mary while only Mary gains true, justified, cognitively significant phenomenal knowledge. Saying that a description predicts their words without giving us Mary's standpoint does not answer that question.

#### The horn the book takes

The house account takes the second horn. Its C consists in a topic-neutral physical architecture: an occurring world-presenting state anchors a token demonstrative and enables later recognition of that state-kind. A full physical duplicate has the same architecture. C explains the conceptual isolation, recognitional dispositions, quasi-phenomenal judgments, and gap-talk shared by the duplicate and us. C alone does not explain the whole of E.[36]

The remaining difference concerns what the concepts refer to and whether the beliefs formed with them are true and justified. Here the identity claim bears the weight. If phenomenal character consists in the physical world-presenting profile argued for in Chapter 6, Mary's acquaintance-grounded concept reaches that physical state; its exercise can yield true, cognitively significant knowledge. A zombie stipulated to lack phenomenal character cannot share that truth while lacking the very state to which the concept refers. Chalmers can redescribe the zombie's corresponding concept and belief, but the difference in E then follows from the disputed identity and its denial in the zombie scenario.

This reply accepts Chalmers's central limitation. The phenomenal-concept strategy does not support the identity, refute the zombie argument, or vindicate type-B physicalism independently. It elaborates the epistemology of a type-B view under the assumption that the identity has been earned elsewhere. Chapter 6 supplies the comparative identity case. The present mechanism explains why that identity leaves a permanent epistemic gap for creatures with our architecture. No circle gets disguised as a victory: C explains the route; identity fixes the referent and truthmaker.

The published physicalist replies often narrow the explanandum in related ways. Carruthers and Veillet distinguish a thick phenomenal reading of E from a thin topic-neutral one. Díaz-León argues that a suitable reading lets C remain physically explicable and explanatorily adequate. Balog disputes Chalmers's treatment of substantial phenomenal belief. Those replies matter, but the house position needs none to make C explain more than it does. It takes the second horn and posts the limit.

**Goff's opacity/translucency challenge**

Philip Goff presses a different objection tailored to a posteriori physicalism.[37] He divides concepts into four groups. A *transparent* concept reveals the nature of its referent. A *translucent* concept reveals part of that nature. A *mildly opaque* concept reveals only contingent features that identify the referent in the actual world. A *radically opaque* concept reveals neither essential nor accidental identifying features.

Goff's premise remains modest. Thinking of pain in terms of how it feels reveals at least something of what it is to feel pain. Full transparency need not hold; translucency suffices. If the type-B physicalist makes the phenomenal concept radically opaque, first-person thought seems no richer than *the property Kev is thinking about now*. Reference has been secured, but the thinker has learned nothing of the property's nature. If the concept is translucent, the physicalist owes an intelligible account of what gets revealed. She cannot say that acquaintance reveals a phenomenal aspect alongside an unrevealed physical aspect without conceding property dualism. Nor can she say that it reveals part of one wholly physical property while insisting that nothing physical is revealed.

The house answer accepts limited translucency. Acquaintance reveals part of the state's constitutive relational profile: what gets presented, the mode of presentation, and some of the determinacy, field structure, attention, bodily address, and evaluative force organizing the episode. It does not reveal every relation in that profile, including the content-fixing history, nor does it reveal which

neural dynamics realize the profile. Ignorance of the realizer is not ignorance of a hidden part of phenomenal essence; property and realizer must remain distinct.

Goff argues that if the revealed aspect belongs to a wholly physical property, the phenomenal concept must reveal something physical. The book agrees *de re* and denies the further *de dicto* step. On the independently defended identity, the disclosed relational profile is physical, although acquaintance does not classify it under the concept *physical*. Seeing a red surface and knowing that one sees a physical property need not arrive through the same concept.

This commits the view to a limited version of Goff's Thesis of Dubious Intelligibility, so its intelligibility needs an example rather than a label. Sight and touch can each disclose roundness through distinct, noninterdefinable recognitional capacities; neither route supplies a Euclidean analysis, yet both reveal something of one spatial feature. The acquaintance and public routes likewise disclose overlapping aspects of one relational profile through different capacities. Their co-reference comes from Chapter 6′s identity argument, not from introspection. The example makes the two-route structure intelligible; it does not turn that structure into independent evidence for identity.

Present pain supplies the hardest case. The book grants that the thin demand *pain occurs now* is perfectly satisfied when the subject attends to the pain. That certainty secures the token and its manifest bodily-evaluative profile. It does not show that the concept reveals every constitutive relation, a nonrepresentational constituent, or the complete metaphysical nature of pain.

The resulting view is revisionary in one exact respect. First-person acquaintance reveals more than bare reference and less than complete essence. It preserves the felt datum and its practical significance while denying that the datum announces all of its metaphysics. Acquaintance explains the route. The identity claim still does the metaphysical work.

### B.9 Swampman and the History Requirement

*(Deep-dive from Chapter 15, Teleosemantics.)* Section 15.6 states the verdict in brief: at $t = 0$ a historyless duplicate can possess a practical stake while lacking representational content; under Chapter 6′s identity claim, he therefore lacks phenomenal character as well. His own history can begin immediately. The full objection-and-reply lives here: Davidson's case, the Millikan–Dretske bullet, Tye's dilemma, and Block's phenomenal-externalist pressure.

Suppose history fixes content: a state means *fly*, or *red*, because of what its producer-consumer system was selected and shaped to track. Then picture a creature with no history whatever. Donald Davidson asked us to imagine lightning striking a dead tree in a swamp and assembling, by accident, a molecule-for-molecule duplicate of a living human being. Swampman has your visual cortex to the last synapse. He hauls himself out of the muck, turns toward a ripe tomato, and his eyes do exactly what yours do. Surely, the objection runs, a perfect copy of you sees the red. If content must be earned through history, teleosemantics seems forced to say that he does not.[38]

The book accepts the price. Swampman's present organization gives him a stake: he is a precarious, self-maintaining system for whom some trajectories preserve the organization and others destroy it, and he regulates himself with respect to those boundaries. That may ground autonomy

and practical interest. It does not supply a semantic address. Nothing in his nonexistent history has selected those visual states for tracking tomatoes, redness, food, or any other candidate rather than its reliable companions. The distinction developed in Chapter 15 therefore matters at its hardest point: the stake explains how success and failure may be good or bad *for* the organized individual; producer-consumer history fixes *what* a state is correct or incorrect about. At $t = 0$ Swampman has the first and lacks the second.[39]

The wide identity makes the result more severe. If phenomenal character consists in historically fixed perceptual content, a duplicate with no such content has no phenomenal character. At the first instant there is nothing it is like to be Swampman. This remains the book's verdict. A structural representationalist can instead preserve his experience through his present organization, without any mental paint. Both duplicates could sort the red tile alike; the behavior would not independently select either verdict. The missing history concerns the property layer itself, not just which tomato supplies the singular object. Chapter 6 therefore gives this extension less warrant than its account of ordinary perceptual contrasts. The historical consequence is retained, not independently confirmed by the thought experiment.

The verdict concerns the instant, not the future. Swampman's first perception-guided action begins a learning and consumer history. As his organization recruits, corrects, and stabilizes states in encounters with tomatoes, food, danger, and other particulars, that history can establish determinate correctness conditions. The present stake makes the consequences matter to him; it does not create the content. So the account does not require ancestry or carbon. It requires a content-fixing history and producer-consumer organization.

The comparison with artificial systems now cuts differently. Swampman begins with autonomous organization and no content-fixing history. A trained model begins with a rich causal history, learned sensitivities, and task-shaped producer-consumer structure that may support genuine mechanism-level content. The model may still lack determinate distal reference, whole-system understanding, autonomy, or consciousness; none follows merely from granting content at one grain. Swampman's history can begin through consequential engagement. The machine case turns on what its training and corpus-mediated history fix, how content-bearing mechanisms integrate across the larger system, and what additional organization supports autonomy or experience. The absences are different, and neither requires a proprietor to mint meaning.

### B.10 Tye's Sharpness Argument and the Concept/Property Dispute

*(Deep-dive from Chapter 10, The Case Against Panpsychism.)* Section 10.4 states the book's verdict in brief: the felt or conceivable sharpness of consciousness does not by itself establish a sharp joint in nature. The full dispute lives here — the two-claims anatomy of the inference, the demonstrative-concept diagnosis, the exchange with Michael Antony, and Tye's insistence that sharpness belongs to the property rather than merely to our way of conceiving it.

Tye's late view needs stating in its developed form before the book turns on it. In the 2021 account, sharp fundamental consciousness* does the panpsychist grounding work, while determinate phenomenal character remains tied to representational content and ordinary consciousness can inherit vagueness from its functional organization. The 2024 restatement presses the essential

sharpness of phenomenal consciousness more directly and pairs consciousness* plus content with a transfer proposal for complex subjects. The formulations differ, but the recurring pressure is clear: our inability to frame a genuinely borderline phenomenal case is treated as evidence that the phenomenal base itself cannot be vague.

Set the two claims the book disputes side by side. First: we cannot imagine a borderline case of phenomenal consciousness. Second: phenomenal consciousness, or the consciousness* that grounds it, *is sharp* in nature. The first reports something about what our phenomenal concepts permit us to frame; the second concerns the grain of the property itself. Tye argues that the first reveals the second. The book denies that entitlement rather than denying the datum.

The structure resembles the conceivability arguments already examined through Mary and the zombie: an epistemic premise about what we can frame is asked to support an ontological conclusion about what nature can contain. The unimaginability is real. What remains disputed is whether it reveals the property's grain or the limits of the concept with which we approach it.

The book's alternative starts from how phenomenal concepts operate. We deploy one from the inside, by having the state and attending to it. It does not work as a description with adjustable conditions of application; it behaves more like a demonstrative — *this*, the thing I am now undergoing. You cannot easily form the thought *there is no fact of the matter whether I am now undergoing this* from the inside, because the having of the state supplies the basis of the concept's application. The concept is all-or-nothing in use. That need not settle the grain of what it tracks: a thermometer that reads only "hot" and "cold" would not show that temperature comes in exactly two sharp kinds. Michael Antony presses this line — a sharp concept may track a vague worldly kind — and Tye refuses it, insisting that sharpness belongs to consciousness itself rather than merely to the concept.[40] The argument therefore reaches a genuine disagreement over what first-person conceivability reveals, not a verbal misunderstanding.

If the book's diagnosis is right, phenomenal consciousness can consist in vague physical-representational goings-on even while recognitional concepts present their application as all or nothing. Panpsychism then loses one of its principal motivations; it has not been refuted by conceptual analysis. If Tye is right that sharpness belongs to the property's essence, the book's vague physical threshold cannot be the whole story. That is the live fork. Chapter 10 prefers the first branch because it fits the identity case already developed and avoids a fundamental phenomenal base plus transfer law; the preference remains comparative and defeasible.

### B.11 Functionalism's First Two Replies

*(Deep-dive from Chapter 20, Functionalism and Computationalism.)* Section 20.4 meets the first two replies to the homunculus-nation in brief and points here for the full hearing: the bullet-biter who accepts that the nation feels, and the dilemma that world-involving content is either mere relabeling of the functionalist's gap or closet dualism. (Objection 3, Chalmers's organizational-invariance argument, stays in the chapter as the reply to beat.)

#### Objection 1: Bite the Bullet

The orthodox functionalist does not flinch. *The nation feels pain* (yes, however bizarre that sounds) *and your refusal to believe it is not an argument but a prejudice.* Our intuitions about who can

feel were trained on creatures with faces, fur, and eyes; they are parochial witnesses, hopeless judges of radio networks and continental populations. We once found it obvious that the sun went round the earth. The strangeness of the homunculus-nation is a fact about the narrowness of our imaginations, not about the distribution of consciousness. Bite down, and the case loses its teeth.[41]

**Reply to Objection 1.** There is something to this, and I want to grant it before I deny it: intuition-pumping over exotic systems proves little on its own, and a reader who found the whole case resting on *"but it just seems weird"* should walk away. The bullet-biter commits no fallacy and incurs no special obligation to name a mechanism beyond the complete functional organization already specified. If functional role exhausts the mental, the nation has pain; that is the view, not a refusal to give one. The objection earns weight only inside the book's cumulative comparison. Chapters 5–16 argue independently that phenomenal character is world-involving content and that content depends on more than an internally specified role. Against that background, the nation case exposes a comparative cost: the functionalist must attribute a unified felt point of view where the stipulated internal organization supplies no independent content-fixing history for the assembled whole. The book's theory explains why that attribution remains unearned; the functionalist accepts it. No single intuition settles the contest. The question is which package explains more while asking us to swallow less. The nation, in short, is diagnostic rather than demonstrative: it exposes the sufficiency thesis's price to a reader who already holds Chapters 5–16, and its whole weight is the weight of that accumulated case. The positive ground is the one the intuition never supplies: the internally specified organization leaves unsettled whether the nation's states possess the complete world-involving, mode-sensitive, perceptually organized, and poised structure the identity claim requires. Giving the assembled whole an appropriate content-fixing history and perceptual organization could change that evidence; the bare wiring stipulation cannot. This is the same bounded conclusion Chapter 20 prices, in the same order.

**Objection 2: You Have Only Renamed the Mystery**

A subtler critic grants that the nation's states are not about anything and presses a different worry. *Fine — but "world-involving content" is just a new label on the old gap. You say the wiring lacks aboutness; I say aboutness is precisely what we do not understand, and you have explained the obscure by the more obscure. Worse: you claim the missing ingredient is not a non-physical extra, yet "genuine meaning" looks suspiciously like the very thing the dualist was gesturing at, now wearing a naturalist's coat. Either grounding is more functional organization (in which case you are still a functionalist and owe the nation its pain) or it is something extra, in which case welcome back to dualism.*

**Reply to Objection 2.** This dilemma is the sharpest thing in the neighborhood, and the answer to it carries the case, so I will be careful. Take the second horn first. Is world-involving content "something extra" in the dualist's sense? No — and the difference is not a dodge. The dualist's extra is *intrinsic*: an inner quality, a private glow, something that could in principle be present or absent with every physical and relational fact held fixed. Grounding is the opposite kind of thing. It is *relational and wholly physical*: a matter of which worldly features have, over a causal history, fixed which states and guided which consumers. Nothing immaterial is added; what is added is

a *world*, and the system's tracked traffic with it. You can be the most thoroughgoing physicalist alive and insist on grounding, because grounding is made entirely of physical relations — just relations the functionalist declined to count.

Now the first horn, which is the one with real bite. *Isn't grounding just more function — long-armed function, reaching out to include the causal links to the world?* And here I concede the words while denying the victory, because the concession is fatal to the thesis under examination, not to mine. Yes: you may, if you like, call world-involving content a species of "wide" or "long-arm" functional role, the functional story enlarged until it takes in the states' causal-historical bonds to the very things they are about.[42] Call it that. But notice what has happened. Machine functionalism, the thesis Chapter 20 set out to test, the thesis that makes the strong-AI claim live, individuates mental states by their *internal* role: inputs, outputs, and relations among inner states, the whole specification readable off the wiring diagram with the world switched off. That thesis is the one Block's nation satisfies to the letter, and it is the one that fails. The moment you reach past the diagram and let the *world* in to fix content, the moment two internally identical systems can differ in what their states mean because they are plugged into different environments with different histories, you have conceded the precise point at issue: organization *alone*, the wiring with the world bracketed, does not suffice. Whether we then file the enriched view under "wide functionalism" or under "world-involving representationalism" is a question of shelving, not of substance. The wiring diagram, read as the functionalist reads it, is no longer the whole story. That is all Chapter 20 needs, and the relabeling hands it over.

### B.12 The Stake: Sources of a Term

*(Background to a term the book leans on from Part Two through Part Four.)*

A system's *stake* in the world — the continued realization of its organization riding on its own regulation under environmental pressure — names a condition with a settled intellectual pedigree, not a coinage of mine. Gathering the sources in one place shows the word carrying recognized work rather than smuggling in a metaphor. Whether any state thereby answers *to* the world remains a further question.

I take precarious adaptive closure as the best current account of autonomy and practical stake, not as the only possible account: a better physical explanation could replace it while leaving questions of content, understanding, and consciousness to their own evidence.

The taproot is Hans Jonas. In *The Phenomenon of Life* (1966) the organism's existence arrives not as a possession but as a task: metabolism holds a living form together only by continually spending and replacing the very matter that composes it, so that the form persists precisely by never resting in any of it. Jonas named the result "needful freedom" — a freedom from any particular parcel of matter bought at the price of needing matter as such. Existence that must be earned, moment by moment, against dissolution gives the first and deepest form of having something at stake.

The enactivist tradition supplied the mechanism and the standard vocabulary. Building on the autopoiesis of Humberto Maturana and Francisco Varela, Ezequiel Di Paolo separates a system that merely holds itself together from one that also registers how it fares and acts on the reading. A living system counts as *adaptive* when it regulates itself with respect to the boundaries of its own

viability ("Autopoiesis, Adaptivity, Teleology, Agency," *Phenomenology and the Cognitive Sciences* 4 (2005): 429–452, sec. 3, where the definition appears in display form). Such regulation responds differentially to trajectories toward and away from the viability boundary. It counts as *precarious* when its constituent processes, taken in isolation, would run down or extinguish in the absence of the organization they sustain — so that those processes stand not merely conditioned by the network but enabled by it. The precariousness condition receives its technical definition four years later, in "Extended Life," *Topoi* 28 (2009): 9–21, sec. 4, where Di Paolo introduces it precisely to keep operational closure from going trivial; the 2005 paper uses the word only informally, in the course of reading Jonas. Evan Thompson calls the wider activity "sense-making" (*Mind in Life*, 2007); the narrower claim here concerns only the functional standard that regulation establishes.

John Haugeland fixed one proposed bearing on intentionality. Understanding, he argued, requires that a system *give a damn* — that it care about itself and its situation — so that a device performing formal operations while caring about nothing understands nothing, however fluent the performance ("Understanding Natural Language," *Journal of Philosophy* 76, no. 11 (1979): 619–632; reprinted in *Having Thought: Essays in the Metaphysics of Mind*, 1998). His formulation remains a serious challenge to accounts that reduce understanding to formal performance. This book now treats caring as a possible enrichment of autonomous understanding, not as a demonstrated prerequisite for content or cognition.

Brian Cantwell Smith carried a related requirement into the age of machine learning. His distinction between *reckoning*, prodigious ungrounded computation, and *judgment*, the answerable and world-committed kind, rests on an "existential commitment" that present systems lack (*The Promise of Artificial Intelligence: Reckoning and Judgment*, 2019). The distinction keeps pressure on any account that infers judgment from performance. Whether existential commitment proves necessary remains part of the dispute rather than a premise the book can simply collect.

Where a term more technical than *stake* is wanted, *precariousness* names the condition. *Normative precariousness* needs more care. It can name a system whose own organization supplies a standard of functional success and failure; it should not suggest that bare existence generates a categorical ought.

### The causal-success objection

Luca Corti has pressed the point directly. Organizational closure, he argues, can explain the causal conditions under which a system exists, but an explanation of why a system persists does not entail that it *ought* to persist or that any component *ought* to perform its role ("Organizational Normativity and Teleology: A Critique," *Synthese* 202 (2023): 96). Put bluntly: a hurricane also has conditions of existence. Describing what would make it dissipate does not give the weather obligations. If a hurricane displayed precarious adaptive closure, it too would earn the thin functional standard defended here; ordinary persistence alone earns nothing.

The objection defeats a stronger claim than this book needs. The stake does not derive categorical normativity from viability. It supplies a theoretical standard of functioning with three features: the standard operates in the system's own regulation; departures alter that regulation before the system disappears; and successful regulation helps maintain the organization that produces

and sustains the regulator. These remain descriptive facts. Nicholas Shea makes the same deflationary point about teleosemantics: correct and incorrect representation can form a descriptive distinction to which stronger norms may later apply; biological well-functioning need not carry an irreducible ought (*Representation in Cognitive Science*, 2018, sec. 6.5).

The term *norm* therefore marks an organization-relative functional standard, not a reason that binds a rational agent. Readers who reserve *normativity* for the latter can substitute *internal functional standard* throughout. What matters is that the organization differentiates trajectories in its own activity, and the consequences return to the organization whose activity produced them. Viability may ground practical interests. It does not supply a representational satisfaction condition. Content depends instead on a history and producer-consumer organization that privilege one accuracy condition over its rivals.

**Where the system ends**

That last sentence raises the boundary problem. Why treat a frog or robot as the system and its pond, server, or data center as a platform? Skin, legal ownership, and the engineer's block diagram settle nothing. The present account locates the boundary through *closure of constraints*. Maël Montévil and Matteo Mossio define the relevant organization through reciprocal dependence: each constitutive constraint contributes to maintaining others that, in turn, maintain it ("Biological Organisation as Closure of Constraints," *Journal of Theoretical Biology* 372 (2015): 179–191). External conditions can remain causally indispensable without entering the closure. Oxygen enables a mammal; the mammal does not maintain atmospheric oxygen as one of its constitutive constraints. A cloud scheduler may preserve a model process; if the model's activity does not help maintain the scheduler or the realizing machinery, the dependence remains one-way. Montévil and Mossio's apparatus adds the operational mark the hard cases need: a constitutive constraint is conserved at the timescale of the process it harnesses and is itself maintained, on a slower loop, by what the organization does. The scheduler fails from both directions at once — nothing the model does maintains the scheduler on any timescale, and the scheduler's persistence owes nothing to how well the model tracks. Fuel, oxygen, and rented compute are consumed or supplied across the boundary; constraints are *held in place* by the loop. That is the difference between what a system runs on and what it is.

The test therefore asks which constraints belong to the reciprocal maintenance network over the time scale on which the candidate organization persists. Include them even when they cross a chassis or a building. Exclude merely enabling conditions, even when their removal would kill the system. The result should remain stable under modest changes of descriptive grain; if a small redescription flips every relation from constitutive to enabling, the evidence does not yet fix a boundary. If an artificial agent acquired energy, repaired its hardware, allocated compute, and preserved the infrastructure that in turn preserved it, the larger distributed organization could constitute an autonomous individual with a practical stake. Moving the boundary outward would then change the verdict because the causal organization changed, not because the analyst grew generous.

Closure can also nest. Cells maintain themselves inside organisms; organisms rely on symbionts; social systems recruit already minded members. Montévil and Mossio treat such cases as hier-

archical closures rather than demanding one metaphysically privileged seam. This supplies a disciplined test for autonomy, not a semantic allocation rule. The level at which content supports belief, understanding, or action depends on a further pattern of consumer use, correction, memory, inference, planning, and control. Location and causal dependence do not confer whole-system attribution. The same signal can enter a cell's economy in one use and the organism's in another.

The social case deserves the same discipline. A company can depend on its workers and the workers on the company without their minds automatically composing one further subject. Economic reciprocity alone is not enough, because reciprocal services may still connect independently maintained agents. Look separately for the closure relations that bear on organizational autonomy and for system-level cognitive operation: error detection that changes organization-wide policy, conflict resolution among competing processes, resource reallocation, and learning that changes how those processes work across turnover in individual members. A firm, colony, automated factory, or distributed robot may therefore constitute an autonomous organization, contain genuine content-bearing mechanisms, or both. Neither achievement yet establishes an understanding or conscious subject; that later question asks whether content enters a flexible cognitive economy capable of integrated inference, planning, and revision. The result may be plural, graded, or empirically uncertain, but it is not arbitrary: reciprocal maintenance bears on autonomy, while integrated consumer work bears on whole-system cognitive attribution. Section 15.6 states the strict etiological limit: present organization without selection or learning history supplies no representational content. On the system/platform application, see Chapter 24.

### B.13 The Ladder of Distinctions

*(Deep-dive from Chapter 24, The Conscious Robot.)*

The charge that the constructive case runs *too quick* almost always comes of holding two or three of these notions in a single hand. Several crowd the same neighborhood, but they do not form one ladder and each has one job.

*Persistence* comes first and settles nothing. A rock persists; a file may persist longer. *Self-maintenance* adds activity directed toward continuation. *Thermodynamic self-maintenance* raises the bar: the system's activity helps constitute the continued existence of the physical organization doing the activity — and the candle flame, the textbook dissipative structure, clears even this rung, melting the wax that feeds it and driving the convection that supplies its oxygen. In organisms, metabolism performs this work. The general condition does not require life or carbon. *Adaptive regulation* adds the ability to register movement toward and away from viable conditions and vary the response. Here success and failure first become possibilities enforced by the system's own dynamics.

The remaining dimensions do different jobs. *Practical stake* arrives where adaptive regulation helps preserve the system's own organization; this book's current working account may be replaced without changing the autonomy and practical-interest questions it serves. *Address* arises through selection, learning, tracking, and producer-consumer use that establish what a state is correct about. *Whole-system understanding* requires those contents to enter a flexible economy of

memory, correction, inference, learning, planning, and action. *Poising* and perceptual organization mark the further condition the identity claim associates with consciousness. The separations block illicit promotions: persistence is not self-maintenance; self-maintenance is not adaptivity; stake is not content; content in a mechanism is not whole-system understanding; understanding is not yet consciousness.[43]

Run three cases. A candle preserves its flame-pattern through a stream of fuel the flame itself liberates — melting the wax below it, driving the convection that feeds it air — a genuine thermodynamic self-maintainer, and conceding the flame this much is no embarrassment: it shows why adaptive regulation must be considered separately. The flame cannot register movement toward its own extinction and vary its response; it burns identically toward the guttering end. A starfish mends itself with matter and energy it acquires and organizes, the repair serving the continuation of the very organism doing the repair. It clears the thermodynamic and adaptive floors, giving it a practical stake; its evolutionary and individual history separately supplies addresses for its states. A conventional virtual-machine process can react to a represented health variable with great sophistication while the host platform maintains the physical process independently of that success. Training may still give the health state genuine content, while the architecture supplies little evidence of autonomous self-maintenance. Remove the backup and nothing important changes. Couple the process's activity to energy acquisition, repair of the realizing machinery, and preservation of the boundary on which that machinery depends, and its autonomy changes even if a copy can still be made. The distinctions ask what the system does, not what the administrator can copy.

### B.14 Why Restorability Is the Wrong Criterion

*(Deep-dive from Chapter 24, The Conscious Robot.)*

The abandoned criterion — this book's own earlier line, given up in the open rather than quietly rewritten — had a clean dramatic shape: a genuine understander would be something you could no longer back up. The shape was cleaner than the argument. Three objections expose why.

First, consider two physically and organizationally identical systems. An administrator has saved a snapshot of one and neglected to save the other. If the file changes the systems' semantic standing, content depends on an event wholly outside both systems. If the *possibility* of a snapshot changes it, the criterion simply redescribes copyability as guilt. Neither route explains why one running system has content, autonomy, or consciousness and the other does not. A backup may reveal the architecture; it cannot constitute the absence.

Second, any finite physical system is copyable in principle at some level of description. Swampman makes the point with unnecessary weather. If perfect duplication voided a stake, better scanning technology would steadily drain autonomy or practical interest from organisms without altering a molecule in them. The absurdity belongs to the criterion, not the scanner. A duplicate raises questions about which individual continues and what history the newcomer inherits. It does not reach backward and revise the original's regulatory organization or content.

Third, take the identity question seriously. Perhaps a restored process is the same agent resumed. Perhaps it is a successor with inherited memories. Perhaps, with Parfit, identity is not what

matters. Even the friendliest horn changes nothing: grant that continuity suffices and the restored process counts as the same agent resumed — what it inherits is form and record, and whether the running system's activity now maintains the organization realizing it remains exactly as open as before the copy. Every horn leaves the substantive questions standing. Does adaptive activity help maintain the physical organization realizing the system? What history fixes the addresses of its states? How are those states integrated and used? A theory of copying cannot answer any of them.

The old criterion did contain one good observation. Current software agents are restorable because their persistence usually resides in infrastructure outside their regulatory loop. The same platform keeps running whether the agent tracks well or badly; the represented danger and the physical continuation come apart. That explains why a checkpoint often serves as useful evidence. Remove the checkpoint and the architecture remains. Redesign the loop so the agent's activity acquires energy, repairs the realizing machinery, and maintains the boundary on which that machinery depends, and the architecture changes even if a copy can still be made.

The replacement is therefore plural: on this working account, thermodynamic self-maintenance coupled to adaptive regulation bears on autonomy and practical stake; world-involving history and producer-consumer use fix content; integration bears on whole-system understanding; perceptual organization and poising bear on consciousness. Copyability concerns identity and succession. These debates may meet in an actual machine, but they do not share a premise.

### B.15 Teleosemantics: The Technical Rounds

*(Deep-dive from Chapter 15, Teleosemantics.)* Section 15.2 states the verdict in brief — the systematic case shows function necessary but not sufficient — and Section 15.3 keeps Fodor's coarseness objection in compact form. The full machinery lives here: Mendelovici's reliable-misrepresentation argument worked through color, and Fodor's asymmetric-dependence round.

This picture has a problem, and the problem has a name. Angela Mendelovici named it in a 2013 paper, careful and unkind in equal measure: *reliable misrepresentation*.[44] Her argument is not the frog argument. She is explicit that the disjunction problem (the problem of explaining how the frog's state can mean *fly* rather than *fly-or-BB*) is not the problem she means to raise. Her argument requires the misrepresentation to be *systematic*, not occasional. To see what that requires, consider color.

Consider the color-error hypothesis Mendelovici uses to put pressure on tracking theories. Suppose experience presents a ripe tomato as having a vivid, simple property that no surface has. A state that reliably tracks reflectance might still misrepresent that further property. The book rejects this as an account of ordinary color: it defends reflectance-type realism. But rejecting the hypothesis differs from showing that a theory of content leaves room for it. The test concerns whether the theory can accommodate systematic error without defining it away.

Hold the example loosely. The argument does not depend on settling whether colors *are* surface reflectances, dispositions, or eliminated altogether. Nor does it depend on the contested assumption I just flagged — that color phenomenology presents redness as intrinsic rather than relational. That assumption names the Galilean intuition that drives a great deal of philosophy of mind, and

direct realists deny it. The argument depends only on the broader possibility (which Mendelovici argues is more than a possibility) that some of our perceptual representations run *systematically* off in this way. Not occasional slip-ups. Not exotic illusions. Standing features of how the system operates.

Now ask what teleosemantics has to say about the matter. The teleological story tells us that a representational state means whatever its evolved function says it means. Its function gets fixed by what the state was selected *for* in the relevant ancestral environment. So what was your color-representing apparatus selected for? Presumably, it was selected because creatures whose visual systems produced these states reproduced more successfully than creatures whose visual systems did not — by, for example, discriminating ripe fruit from unripe, or warning coloration from camouflage, or kin from non-kin. The function is to discriminate, to categorize, to enable adaptive behavior.

Here teleosemantics walks into the wall. If the function of your color-representing state is to *discriminate certain surface properties under certain lighting conditions*, then your state means exactly that. It does not misrepresent the tomato at all. It accurately represents the tomato as having the relevant discriminable surface property, which the tomato in fact has. The vivid, intrinsic-seeming redness drops out of the content. Asked what the state represents, the teleo machinery hands back an answer that makes systematic misrepresentation *practically* unavailable: the state means whatever feature was tracked reliably enough to be selected for, and reliable tracking is part of what fixes the content. Mendelovici puts the modal claim carefully: tracking theories make reliable misrepresentation "practically impossible," and that counts as a major consideration against them. The point is not that no single coherent description of reliable misrepresentation can be cooked up within the framework. The point is that the framework's resources for handling such cases fall systematically short.

What does the experience attribute? Mendelovici asks us to leave room for content that goes beyond what tracking secures. The reflectance realist denies that ordinary color has that content: a property can look simple without being presented as physically simple. The identity thesis needs an account of the actual property-content, whichever view wins. Tye defends the surface-color view.[45] That defense does not settle whether a theory of content can allow the kind of error Mendelovici describes.

A defender of teleosemantics has replies, and the literature since Mendelovici has not stood still. Karen Neander has developed an "informational" teleosemantics that tries to fix the indeterminacy worry by appeal to natural-factive information and second-order similarity relations.[46] Peter Godfrey-Smith, a friend of the program, has stepped back to ask whether it has stayed in proper contact with the phenomena it was supposed to explain.[47] His diagnosis treats indeterminacy as a deep structural worry rather than a local puzzle. None of these moves dissolves Mendelovici's problem. They redistribute it. The teleo theorist can tighten the function, in which case content threatens to track only what the system reliably picks up, leaving the phenomenology unexplained. Or the teleo theorist can broaden the function, in which case content turns indeterminate, since several incompatible candidates each fit the historical record. The tradeoff holds.

I take the right lesson from this to be modest. The reliable-misrepresentation argument does not show that teleosemantics is false. It shows that teleosemantics cannot be the *whole* story about content. Evolution can do a great deal of work in explaining why our representational states exist, why they have the functional profiles they do, and why misrepresentation in principle remains possible (a state that performs its function only sometimes obviously can fail to perform it). What evolution cannot do, by itself, is fix the content finely enough to allow systematic mismatch between how things seem and how things are. For that, we need a story about how representational vehicles *carry* content beyond what natural selection has ratified — a story that lets the visual system represent the tomato as red even when nothing red, in the relevant thick sense, is on offer.

The consumer side gives us a place to look, but not a completed answer. Dretske distinguishes what a state now indicates from what it has the function of indicating.[48] That distinction allows present correlation to come apart from content. It does not by itself explain how a state could have acquired content concerning a thick color property that was never instantiated.

Take the proposed consumer explanation. A perceiver matches a strawberry to a tomato across changes of light, groups their surfaces together, and corrects mismatches. Those responses and their learning history can constrain a candidate mapping to reflectance-types. Now suppose the theorist says that the consumers treat both surfaces as having *thick redness*. What fact, described without that phrase, selects the thick property rather than the reflectance-type? The matching history just given does not choose. Calling it use *as thick red* imports the content it was meant to explain.

The book retains reflectance realism and the role of history in fixing property-content. The perceptual comparisons support that program; they do not supply a general answer to Mendelovici's challenge about systematic error. Nor can anxiety supply an easy substitute for color. A harmless morning that feels menacing illustrates a candidate mismatch only if an independent account first fixes what the mood presents. Chapter 7 and Section 15.3.2 have not completed that account.

Mendelovici argues that tracking theories make reliable misrepresentation practically unavailable, especially in clean cases without a further defect in functioning. The objection need not claim that every case is logically impossible. Historical consumer use helps distinguish correctness from present success; showing that it also accommodates her clean systematic cases remains further work. The theory cannot borrow a solved semantics of feeling to claim that result.

Jerry Fodor saw the difficulty for teleosemantics earlier than its friends did, and pressed it into the strongest standing rival the naturalistic program has.[49] Fodor spent a career arguing that the mind runs on a language of thought; he put the point with the frog. Suppose the frog's detector has the biological function of getting the frog fed. A frog whose inner state means *fly* and a frog whose inner state means *fly-or-BB-or-any-small-dark-moving-speck* are, in the ancestral pond where the only small dark moving specks were flies, as fit as each other. They catch the same things, leave the same number of descendants, fare identically under selection. Biological function, being a matter of what got the lineage fed, simply cannot tell the two contents apart, because the fitness comes out the same whichever way you read the state. This bites deeper than the BB that orthodoxy

already handled: no BB haunted the ancestral pond, but the rival content here runs co-extensive with *fly* in the very environment that did the selecting, and equal fitness leaves nothing to choose between them. So function is too coarse to fix the address. That is a real result, and it drove Fodor to build his asymmetric-dependence theory instead: meaning, he proposed, gets fixed not by what a state was selected to track but by a counterfactual pattern of dependence among the things that can cause it, so that the BB tokens *fly* only because flies do, and not the other way around. The teleosemanticist cannot wave this away. Survival underdetermines content as Fodor said.

And the rival does not get off lightly either: Fodor's asymmetric dependence arguably cannot spell out its counterfactuals without reaching for the very semantic and functional facts they were meant to ground.

**Round Three: What fixes *fly*, if anything?**

Granting Fodor's coarseness result does not leave every candidate tied. Selection is a causal process, and causal explanations do not permit free substitution among co-extensive descriptions. The frog survived because it captured nutrition, not because it captured blackness. That weighs against *small dark thing* as an account of the task function even if every nutritive target in the ancestral sample happened to look dark. The point does not yet exclude *fly-or-speck*. A branch cannot lose merely because no specks happened to occur. It loses if specks figure in no Normal condition of consumer success, lack the nomological and counterfactual support carried by the fly branch, and improve no causal explanation of how the producer-consumer system performs its task. A logically available disjunction is not yet an explanatory one, but historical silence alone does not decide the case.

Nicholas Shea's varitel semantics makes the positive method explicit.[50] First identify the robust task function whose performance explains why the mechanism stabilized. Then identify the tracking relation that figures without mediation in the mechanism's performance of that task. Finally place the mechanism among the organism's other capacities. If the frog uses distinct circuits for worms and other edible targets, the correlation with flying insects may explain why this circuit exists and how the wider feeding economy divides its work. *Fly at (x, y, z)* may then earn its grain. If the same circuit and consumer treat every small moving nutritive item alike, *moving prey at (x, y, z)* may be all the content the evidence supports. *Edible moving thing* and *prey* do not lose merely because English speakers find *fly* neater.

This answer is realist without pretending to metaphysical omniscience. Content consists in the satisfaction condition supported by an intervention-stable causal explanation of the producer-consumer system across actual and relevant counterfactual tasks. Where one candidate tracks a difference that changes performance and another hides it, the architecture supplies discriminatory structure; the theorist merely gets it right or wrong. Where two candidates remain co-extensive and equally explanatory across every such difference, no further fact need choose between them. The state has content at the grain its use supports and remains indeterminate below it. The demand that a frog's signal encode a philosopher's preferred natural-kind concept was never part of naturalism's job description.

The BB case now comes out without sleight of hand. If the content is *moving nutritive prey at this location*, the BB token is false because the pellet does not satisfy that condition. If the content is only *small moving object at this location*, the token is accurate and the feeding failure occurs downstream. Which verdict applies depends on the established task structure. Either way, correctness and successful performance remain distinct: a correct state can guide a clumsy strike, and a false state can produce a lucky meal.

**Round Four: Does training supply a selection history?**

Training can supply a selection history. Mallory argues that each round of backpropagation copies and alters a prior set of weights. The update rule preserves some structure while loss guides change. Producers and consumers acquire functions through this process. On his account, even word2vec can contain signs with original content.[51] Training gives the artifact a real causal past; the lack of biological evolution cannot disqualify it. Selected checkpoints belong to that lineage. A deployed copy may inherit selected functions through descent from trained mechanisms, without having selected itself. The history can explain why its internal states perform their roles. The remaining question concerns what those roles make the states answer to.

The semantic proposal concerns a specific grain. Mallory's Section 5.1 identifies learned embeddings as signs concerning word-types and their contexts. Outputs are compared with distributions in the data, and discrepancies reshape producer and consumer. This supplies positive evidence for world-answerability: the target is the actual pattern of word occurrences in the data, not a number to which a reader happens to assign a meaning.[51a] Human constitution of word-types does not make every judgment about their distribution true by convention.

The independence comparison must match selection, correction, and consumer use. Consider two learners with the same token inputs and matching ordinary predictions. One acquired its mapping through errors measured against sampled contexts; the other through errors measured against a schedule of approved predictions that usually agrees with those samples. Hold token identity and the context window fixed. Let the schedule favor *engine* beside *cat* while the samples still favor *pet*. Does the mismatch with the samples retune the learned state even when a score bonus preserves the favorable verdict? Now change the samples while preserving the schedule and follow the effect through the consumer. A distribution-directed history supports one mapping; an approval-directed history supports another. The latter may itself establish original content about approvals. What it does not establish merely by matching ordinary answers is the further content we ascribe about actual word contexts.

Robot B faces the same comparison in Section 24.1. Its leak diagnosis must answer to fluid escape when a service register says *no leak* and a backup pump rescues the payoff. Robot A has learned states, correction, and consumers too. Its stipulated derived leak content cannot validate the test; the contrasting routes must explain the histories and changed responses. If both robots use the same world-mapping correction, both receive the same semantic assessment. Adding a maintenance loop or a proprietor cannot break that tie.

Mallory's direct comparison with actual distributions favors the distribution-directed account. What his analysis does not report is the matched separation of sampled contexts, supplied targets,

and scores just described. The book asks which relation explains the learned embedding's contribution to its consumer under that separation. Historical and mechanistic evidence can answer without a new experiment, and for the algorithm Mallory describes it does: outputs compared with actual contexts reshape producer and consumer, so the distribution-directed relation explains the embedding's contribution. Online correction and several consumers are not extra conditions on all content. The leak story supplies a conditional answer by construction; Mallory's described algorithm supplies one through its documented history. A particular word2vec implementation must still show from its own record that its training and use instantiate that relation. Its engineered loss, lack of autonomy, and ancestry in human language do not count against it. A rival mapping that explains the same history and use equally well would limit the embedding's claim and the robot's alike.

Two other interventions must be kept apart. Coordinating a change of indices, matrices, and word key can preserve every actual word-to-context relation. Changing which word feeds an index without coordinating the rest alters a candidate content-fixing relation; changing the corpus or the context window can alter the target. Neither maneuver shows that an interpreter's reassignment alone governs correctness. Nor does a corrupt feedback episode instantly replace a historically fixed content. Millikan's distinction between a correspondence needed for proper consumer use and accidental success grounds this separation of accuracy from payoff. It does not make the proposed comparison an infallible test of originality.

Finally, a change to the maple outside the window while the corpus stays fixed tests a proposed reference to that tree. It does not test Mallory's content about word distributions. If history and consumer use establish independence at his limited grain, that warrants original content there without settling distal singular reference. The same restriction applies to B: tracking fluid escape does not by itself establish understanding, consciousness, or welfare.

Content alone does not settle whole-system understanding. Do the states let memory, correction, and planning work together? Can they guide inference, settle conflicts, and direct action as the system learns? A monitoring loop added for one task does not establish that wider economy. Flexible use across changed conditions supplies the positive evidence. An automated factory, distributed robot, or trained model could earn this attribution within a bounded domain. The account must accept that result rather than protect a human monopoly. Reciprocal self-maintenance would support autonomy and a practical stake. It would neither turn derived content into original content nor make otherwise isolated states add up to understanding.

The verdict can move in either direction. If an artificial lineage's history and use fix accuracy beyond an assigned verdict, the case for original content strengthens. If our proposed reading adds a standard that the history and use do not support, that reading remains dependent on us. A failed trial alone decides neither: it may expose an error under a standard already fixed. Robots and artificial-life systems offer possible cases, not privileged routes. The context learner must receive the same hearing. Wider use across learning, inference, correction, planning, and action would strengthen the case for understanding. Resource gathering and repair that preserve the system's own organization would support autonomy separately. Manufacture disqualifies none of these achievements. For the algorithm Mallory describes, that history warrants original content about

word-types and their contexts, conditional on the teleosemantic account used here. That still leaves distal singular reference unproved. Whole-system understanding, a subject, consciousness, and welfare would need their own evidence.[51b]

**Round Five: Does the arrow run the other way?**

One rival has stood off to the side through all four rounds, and it disputes none of the answers given here. It disputes the direction.

Angela Mendelovici and David Bourget defend the phenomenal intentionality theory, on which "all actual intentional states, or at least all actual originally intentional states, arise from phenomenal states."[52] Tracking theories, on their reading, get the order backwards: aboutness does not accumulate out of correspondence laid down by selection, it comes from what experience is already like. And the reliable-misrepresentation argument that opened this excursus belongs to that case — Mendelovici offered it as ground-clearing, not as a local repair. I have leaned on her demolition for several pages. Fairness asks that I say what she meant to build on the cleared lot, and why I decline to help her build it.

Two costs, and the first one names an organism.

If original intentionality arises from phenomenal states, then nothing unconscious means anything on its own account. The bacterium's content becomes derived, held on loan from us. This book holds the reverse, and holds it as essential: aboutness and consciousness converge at the top and separate at the bottom. The bacterium swimming up a sugar gradient may carry content with nobody home to enjoy it. Concede that ground and Part Three loses its bottom rung — the path from tracking to understanding has nothing left to stand on.

The second cost arrives disguised as a gift, which is exactly why it wants refusing out loud. The phenomenal intentionality theory would settle Part Four in a sentence: no phenomenal states in the language model, therefore no original intentionality, therefore no understanding — the verdict handed over before a single question about architecture gets asked. I would rather not have it on those terms. A conclusion that follows from a premise about consciousness, rather than from what a system does and fails to do, wins the argument by making the argument unnecessary. Content-fixing history and whole-system integration must show their work independently.

Underneath both sits the structural point. The deflation of Chapter 11 needs the identity claim to reduce, and the reduction needs content naturalized first, independently of phenomenal feel. Run the arrow the other way and that independence goes: phenomenality does the explaining, the reduction stops reducing, and Chapter 11 loses the independent content account its proposed deflation needs. Mendelovici and I cannot both have the order right. The order carries the architecture.

### B.16 Illusionism: The Rival That Agrees About the Gap

*(Deep-dive from Chapter 11, The Gap That Isn't There.)* Keith Frankish asks why we judge experience to possess phenomenal properties if physical organization can explain everything we say and do. His strong illusionism retains sensory, motivational, and affective organization. It also permits a functional sense in which experience feels like something. What it denies is that the properties

behind our introspective judgments really possess phenomenal character, whether physical or nonphysical. Frankish calls them *quasi-phenomenal*: they lead us to ascribe a character they do not have.[53] Dennett's earlier attack on ineffable, intrinsic, private, directly apprehensible qualia supplies part of the background without making every version of their positions interchangeable.[54]

The book shares much of that project. An inner witness, a private coat of color, and an infallible introspective reading have no place in its account. Both approaches must explain why the hard problem seems compelling. Chalmers calls this the *meta-problem* of consciousness.[55] Neither theory can win merely by explaining a red/green discrimination, an aversive response, or a report of pain. Both retain the organization behind those events.

Consider a narrower judgment: while looking at a tomato, you judge that it visually looks red to you. This differs from judging that the tomato really is red; lighting could mislead you about that. It also differs from judging that your brain contains an intrinsic red quality. The book holds that the first judgment can correctly recognize your current visual presentation. Its phenomenal reference reaches that presentation itself, not an extra property attached to it and not merely a disposition to make the judgment. Phenomenal character consists in the independently fixed property-content under its visual mode, with its detail, field structure, and attention: one and the same thing under two descriptions. Its occurrence does not depend on your making an introspective judgment about it.

Frankish can preserve the first-order state, its worldly accuracy conditions, and the mechanism that monitors it. None supplies a feature his account simply omits. The disagreement concerns what the judgment correctly attributes. A subject can refer to the right state and still describe it wrongly. The book holds that the judgment can correctly attribute genuine phenomenal character to the world-presenting profile. Strong illusionism denies that attribution while preserving the profile and the judgment. This differs from adding another mechanism or private ingredient. If an illusionist grants the same attribution and reserves "phenomenal" for intrinsic paint, the disagreement at this point becomes verbal. Agreement about the mechanism alone does not make it so.

What supports the affirmative answer? Chapter 6 constrains a visual presentation through discrimination, error, and intervention before using reports of appearance as a test. Suppose changes in focus alter measured spatial detail, and the subject distinguishes those changes while separating them from changes in the print. That pattern supports recognition of a real presenting state. The identity proposal goes further: the character recognized through acquaintance consists in that profile. The recognitional and descriptive concepts differ, though the proposed property does not. Mere competence at reporting the changes supplies no proof that the phenomenal description gets it right.

The lightning analogy shows the room for this proposal. Someone can recognize lightning while holding a false theory of electricity. Correcting that theory need not eliminate what the person recognizes. But the analogy cannot decide whether phenomenal thought works that way. Frankish can accept successful reference to a presentation and locate the error in the genuinely qualitative character attributed to it. He denies even purported physical phenomenal properties, not merely

intrinsic paint. His challenge asks whether the identity account has articulated genuine phenomenality or only redescribed the organization his theory already accepts.

I retain the affirmative account: phenomenal thought can characterize a real physical presentation without recognizing it as physical. The evidence considered here does not settle that account of phenomenal thought against Frankish's. Nor does it establish an economy advantage simply by counting fewer false judgments. That advantage would follow only after we had earned the disputed account of what those judgments attribute. The comparison therefore locates a substantive burden for the identity theory; it supplies no separate victory over illusionism.

François Kammerer presses the danger for any such defense: confidence in the datum may itself be part of the illusion.[56] Mere causal contact cannot answer him. Successful discrimination and correction constrain what the judgment concerns, but can coexist with a false phenomenal description. First-person certainty cannot close that gap in the argument.

Both accounts reject intrinsic paint. Their remaining dispute concerns whether the independently specified presentation genuinely satisfies our phenomenal concepts or merely explains why we apply them. "Pain hurts" states what needs explaining; it does not settle how to explain it.

### B.17 Higher-Order Theories: The Monitor and the Missing Target

*(Deep-dive from Chapter 6, The Identity Claim.)* Chapter 6, note 7b introduces higher-order theories as rivals about state consciousness. David Rosenthal begins with a question every account must answer: why do some mental states occur consciously while others do not? His proposal requires a suitable higher-order thought, a thought to the effect that one occupies the first-order state. Hakwan Lau and Rosenthal develop an empirical case through differences between subjective awareness and objective task performance.[57] A perceptual system can discriminate a stimulus without the subject reporting that she sees it. Explaining the discrimination alone leaves the contrast open.

The higher-order thought need not itself be conscious. It introduces neither an inner eye nor a spectator, and no further conscious thought must watch it. Nor does a first-order account of character automatically refute the proposal. What makes an experience look red rather than green can differ from what makes it conscious at all. A theory could assign the first job to the perceptual state and the second to a higher-order condition. That condition would remain constitutive of state consciousness even if it supplied none of the color.

Misrepresentation sharpens the question without deciding it. Suppose the first-order state presents red while the higher-order thought characterizes it as green-presenting. Or suppose a suitable higher-order thought occurs without the target state. The theory must distinguish the first-order quality, the way the subject experiences herself, and whether the target state exists. Block argues that the targetless case creates an incoherence. Rosenthal replies that a subjective appearance can occur without the state it purports to disclose; he does not say that the higher-order thought brings the absent target into existence.[58] On his account, appearance follows the higher-order content. A variant which instead assigns character to the target must explain how erroneous higher-order content makes that target conscious without fixing its experienced quality. Neither answer becomes incoherent merely by assigning different work to the two levels.

The book rejects a necessary higher-order condition, but the reason must be comparative. Section 6.5.1 proposes a recurrent, selective pattern of two-way influence among attention, memory, inference, planning, and flexible action. Content can occupy that organization without a separate thought certifying that one has it. Higher-order monitoring may help realize the pattern; the proposal does not forbid it. What remains disputed is whether monitoring must occur wherever consciousness does. The identity claim about character cannot settle that separate question.

Lau and Rosenthal's evidence puts the first-order account under real pressure: subjective visibility and confidence need not follow objective discrimination in lockstep. The reply cannot simply declare all those measures equivalent. It must ask whether the complete access-and-control organization, rather than one performance score, stayed fixed. If consciousness changed with higher-order representation while that independently specified organization remained unchanged, the house account would fail. Conversely, a conscious state with the distributed organization but no relevant self-representation would count against a necessary higher-order condition. Those are proposed discriminating cases, not findings already delivered by the transparency exercise.

Uriah Kriegel's self-representationalism brings world-directed and self-directed aspects into one state.[59] That blocks the simplest picture of a wholly separate, targetless monitor. It still owes an account of the relation between those aspects and of why the self-directed one is necessary. The book's preference remains for conscious world-presentation without that additional requirement. Establishing it requires evidence about the organization that sustains experience, not the observation that nobody has found a watcher.

**B.18 Integrated Information Theory: Integration Without a World**

*(Deep-dive from Chapter 20, note 3, and Chapter 24.)* Integrated information theory stands as the most mathematically developed account of consciousness on offer, and the one a reader most reasonably asks this book to face. Giulio Tononi built it from the first-person side outward: begin with what any experience self-evidently is — intrinsic, structured, specific, unified, definite — and demand of any physical substrate that it realize those properties in its causal organization. IIT 4.0 states the axioms and postulates in its current form and analyzes a system's irreducible cause-effect structure through distinctions more exact than the older shorthand of a single Φ score.[60] Consciousness, the theory says, consists in maximally irreducible intrinsic cause-effect structure. Credit what the theory does that its rivals talk about: it treats the first-person datum as a constraint on physics rather than an embarrassment to it and refuses the behaviorist shortcut — the relevant causal organization concerns what a system is, not what it says. Perturbational-complexity measures used in anesthesia and disorders of consciousness belong to an IIT-inspired clinical program; they are not direct measurements or independent validation of metaphysical Φ.

One convergence deserves noting without treating it as confirmation. Under the cited IIT formulation, a purely feed-forward system has **zero** Φ, and Tononi and Christof Koch conclude that even a perfect software simulation of a human brain, run on standard von Neumann hardware, would not be conscious.[61] Part Four reaches caution about present text-centered systems by a different road: their mechanisms may carry inherited or learned content while the evidence for one integrated, perceptually organized, world-presenting subject remains incomplete. The two

views therefore overlap on some architectural suspicions while disagreeing over both the criterion and the evidence.

IIT remains blind to content in exactly the sense Part Two made criterial. Its intrinsic cause-effect analysis concerns internal difference-making however the system stands to a world. In principle a lattice arranged for high irreducibility may receive a rich conscious structure while presenting nothing beyond itself. Aaronson pressed that cost with expander graphs, and Tononi accepted the consequence rather than treating it as a reductio.[62] The unfolding argument adds a different challenge: recurrent and feed-forward systems can duplicate input-output performance while IIT assigns them radically different consciousness. An IIT defender can answer that causal organization, not functional equivalence, is the point; the reply preserves the theory by making the empirical burden explicit rather than dissolving it. The house identity claim runs the order differently. Phenomenal character consists in world-presenting content of the right kind, with integration entering as part of the poising and subject-level organization that make some content one subject's presented world, never as a sufficient condition standing alone. That disagreement concerns theory, not a defect that one formula can settle.

Read as a developed model of intrinsic causal integration, IIT supplies hypotheses Chapter 24′s subject and organization questions can use. Read as a metaphysics — intrinsic cause-effect structure as what consciousness is — it parts from this book exactly where structure parts from world-presenting content. The book can learn from the formalism without treating either the clinical proxies or the feed-forward verdict as borrowed proof.

### B.19 Free Will Without Origination: The Dispute Behind the Capacity View

*(Scholarly material compressed from Chapter 18, The Nature of Free Will.)* The chapter now asks one question: what capacity distinguishes conduct that answers to reasons from conduct in which that capacity has broken down? The surrounding free-will literature is larger. Neuroscientific abolition, Kantian origination, the Consequence Argument, libertarian indeterminism, and unstable folk intuitions all press on the capacity view from different sides. They belong here because none is needed to follow the chapter's bridge from the assembled subject to agency and machines, while each deserves more than a passing dismissal.

#### The neuroscientific abolition argument

Robert Sapolsky and Sam Harris trace decisions upstream through neurons, hormones, upbringing, and the history of a particular body. At no point does an uncaused chooser appear. Benjamin Libet's readiness-potential experiments became the standard exhibit: neural preparation seemed to begin before subjects reported deciding to move.[63]

The evidence establishes that decisions have physical causes. The capacity view grants that before the first electrode is attached. What it denies is the extra premise that freedom requires causelessness. Libet studied arbitrary wrist movements stripped of reasons, and later accumulator models and no-movement trials weakened the inference from readiness potential to an unconscious decision already made. Even the strongest neural prediction would leave the relevant question untouched: can represented considerations enter one persisting economy, alter deliberation, and

govern action? The Parkinson case matters because medication damaged that relation while evaluative judgment survived. It identifies a capacity failure, not a holiday from physics.

### Origination and the ghost's freedom

The older picture asks for more. Kant's transcendental freedom begins a causal series on its own; Ryle's ghost in the machine supplies the durable caricature of a chooser upstream from nature.[64] Popular skepticism concludes that no such chooser exists. Libertarianism seeks an indeterministic opening through which the agent can act. Both accept origination as the prize.

Indeterminism does not secure it. If prior reasons, values, memories, and character do not settle the act, an unexplained remainder has not thereby become authorship. Event-causal libertarians try to place indeterminism inside deliberation rather than beside it, but the luck problem returns: why does this undetermined outcome belong to the agent rather than to chance? Reliable causation looks less like freedom's enemy once randomness becomes the alternative.

Galen Strawson's Basic Argument and Pereboom's hard incompatibilism require separate treatment. Strawson asks for responsibility for the equipment from which one acts and concludes that genuine self-creation is impossible. Pereboom targets *basic desert*, the claim that a wrongdoer deserves suffering simply for the deed. The chapter's reasons-responsive capacity establishes neither self-creation nor basic desert. It establishes answerability: the standing to receive a demand, understand it, and revise.

### The Consequence Argument and identical universes

Peter van Inwagen gives incompatibilism its clean modal form. If determinism holds, present action follows from the remote past and the laws; neither is up to us. David Lewis and later dispositional compatibilists reply that ability concerns what an agent would do under relevant variations in reasons and opportunities, not whether an identical total universe can branch into two futures.[65]

The capacity view should concede van Inwagen's result about the fixed total state. It rejects the further claim that agency requires a second future from that state. Fragility belongs to a glass that never breaks; a pianist retains an ability on an evening she never approaches the instrument. Likewise, a person may possess the capacity to refuse even when actual reasons and character determine acceptance. The counterfactual varies the circumstances under which the capacity would manifest. It does not ask the universe to fork while nothing changes.

Searle presses the same demand through the experienced gap of deliberation. While choosing, the outcome feels open, and deliberation presupposes that openness. But the deliberator cannot read the result from a physical description in advance and must still weigh reasons. That epistemic position makes deliberation causally necessary; it supplies no further metaphysical opening. Remove the weighing or change the reasons and conduct may change. Deliberation works because it forms part of the causal process, not because the process pauses for it.

### What ordinary judgment can and cannot settle

Experiments on folk judgment divide sharply by framing. Abstract deterministic cases elicit incompatibilist answers; concrete wrongdoing often elicits compatibilist ones.[66] The result blocks

any simple appeal to "what we ordinarily mean" by free will. It does not decide which response reveals competence and which reflects distortion.

The excusing conditions offer firmer evidence because they discriminate cases. Coercion, compulsion, delusion, and chemical disruption predict whether warning, argument, reflection, and correction can reach conduct. The power to produce a different future from an identical universe distinguishes none of them. A proposed essence of agency that explains no contrast inside the practice has less claim than the capacity those contrasts track.

**Why the chapter keeps the capacity view**

The reasons-responsive account is deliberately modest. It permits degrees, local breakdown, engineered origins, and borderline cases. It grounds fitness for address, not virtue or punitive desert. Luck should temper confidence about what one deed reveals and what responsibility may fairly exact. Manipulation excuses where it bypasses the agent's weighing, not merely where another hand helped build the starting equipment.

That modesty is also what makes the account useful for machines. Manufacture disqualifies nothing. Fluency proves nothing. The test asks whether represented reasons persist across memory, correction, conflict, planning, and action within one continuing control organization. The excursus records why that capacity is not the ghost's freedom. Chapter 18 asks whether anything actually has it.

### B.20 Assessing Artificial Consciousness Under Uncertainty

This excursus supplies the assessment method behind Chapter 24 and the Coda. It does not turn theories of consciousness into a scorecard. It asks what a named physical system realizes, what each observation bears on, which rival explanation remains, and what would change the verdict.

**The bearer comes first**

"Is the AI conscious?" leaves the subject of the question unfixed. At least five candidates require separate assessment:

1. **The running base model:** model weights, activations, context, and one or more forward passes, without the ordinary product wrapper.
2. **The deployed conversational product:** the model together with system prompts, retrieval, safety layers, session state, and routing used in an ordinary exchange.
3. **A persistent tool-using agent:** a deployed product whose memory, goals, plans, tool calls, and corrections constrain later activity across tasks.
4. **An embodied controller:** a model or agent coupled to sensors and effectors through a continuing perception-action loop.
5. **A wider recurrent system:** any larger organization spanning models, memory stores, tools, sensors, control processes, and perhaps people, but only where those parts participate in one recurrent economy rather than merely supply or repair it.

The book's provisional bearer rule then applies. A candidate subject is the relatively maximal persisting control organization within which contents directly update a shared economy of attention, expectation, memory, goals, correction, planning, and action. Partition tests, memory and

goal interventions, and blocked correction routes locate that economy. They identify a candidate bearer; they do not prove that the organization composes one point of view.

### What survives the functionalist bridge

Butlin and colleagues derive indicators from recurrent processing theory (RPT), global workspace theory (GWT), computational higher-order theories (HOT), predictive processing, attention schema theory (AST), and work on agency and embodiment. Their inference rests partly on computational functionalism as a working hypothesis: if the right computation suffices for consciousness, then a good computational match to a theory's proposed function raises confidence.[67]

This book rejects only the narrow bridge from an abstract or world-bracketed program to consciousness. It accepts material neutrality and realized causal-computational organization. The indicators therefore retain evidential force, with one amendment. Recurrence, broadcast, metacognition, predictive coding, attention control, and agency must occur within the concrete bearer under assessment, and the states they organize must have independently supported content. A functionally elegant but semantically indeterminate diagram may realize access organization without yet realizing the world-presenting content that strong representationalism identifies with phenomenal character. Conversely, world-involving content without recurrence, broadcast, or another supported state-consciousness relation may remain unconscious processing.

The amendment changes weight rather than erasing evidence. It discounts a match confined to a bare forward pass when the claimed bearer requires persistence, and it increases the relevance of an agent or embodied controller whose contents acquire new accuracy conditions through continuing engagement. It also keeps three questions apart: what a state represents, whether it enters consciousness, and which system bears it.

### Mapping the indicator families

The following map records evidential roles. The rows do not add up to a total. Several may arise from one mechanism, and one strong causal result can outweigh several behavioral resemblances.

| Family | What would support it in an artificial bearer | Strongest current deflation | What would move confidence |
|---|---|---|---|
| **RPT** | Algorithmic recurrence within perceptual or input processing that integrates a structured representation and makes later processing depend on the returning signal | Token-by-token iteration, scratchpads, or agent loops may recur only at the product level; a transformer forward pass may remain largely feedforward | Raise confidence with causal interruption of feedback that selectively disrupts integrated, reportable content while sparing comparable feedforward performance; lower it if the same capacities survive removal of the alleged loop |
| **GWT** | Specialized processes competing for a limited workspace whose contents become broadly available for report, reasoning, control, and sequential querying | A shared residual stream or router may support flexible access without encapsulated modules, recurrence, sharp ignition, or phenomenal consciousness | Raise confidence through interventionally verified bottleneck, competition, selective entry, and broad downstream broadcast across independently operating capacities; lower it if "broadcast" reduces to a probe-friendly encoding with no privileged causal role |
| **HOT** | A representation of the system's own first-order representation, used to monitor reliability and update belief or action | Confidence may track input difficulty; self-description may concern a trained assistant persona; a proposed higher-order state may carry only first-order distal content | Raise confidence through cross-task generalization, a self-prediction advantage over matched models, spontaneous monitoring, and interventions that dissociate object-level processing from metacognitive control; lower it when first-order or persona models explain the same effects |
| **Predictive processing** | A hierarchical model of the body, world, and self that uses prediction error to update perception, learning, and action across time | Attention may be weights, not inference; apparent surprise may be a trained output pattern; and a large model may implement useful dynamics without representing its own uncertainty | Raise confidence by [illegible] |
| **AST** | [illegible] | [illegible] | [illegible] |

A July 2026 mechanistic result materially strengthens one row. Wes Gurnee and colleagues identify a small "J-space" of verbalizable representations in tested language models. Causal interventions support report, directed modulation, silent internal reasoning, flexible reuse, selectivity, and broad downstream broadcast.[68] That moves the GWT-like access evidence upward. The authors take no position on phenomenal consciousness, and they mark important mismatches with biological global-workspace proposals: no clean analogues of recurrent loops or encapsulated modules, and no demonstrated sharp ignition. On the present view, the result establishes a privileged access organization inside a running model. Its weight for phenomenal consciousness still depends on the model's content, state-consciousness theory, and bearer.

The higher-order evidence warrants a similar but smaller update. Patrick Butlin finds evidence consistent with limited higher-order representation in language models: calibrated confidence, access to facts about their own behavior, self-simulation, and representations of assistant-persona properties. He also finds the evidence inconclusive. First-order and persona explanations remain live, the distality problem complicates content attribution, and present experiments do not show spontaneous metacognitive control across ordinary tasks.[69]

Goldstein and Kirk-Giannini argue that language agents come close to permissive GWT conditions and could satisfy them with modest modifications.[70] Their argument raises candidature conditional on GWT and on their functionalist interpretation. It does not settle the bearer, the book's wide-content requirement, or the step from access to phenomenality. Chalmers's broader theory-balanced route reinforces the same discipline: recurrence, world and self models, workspace organization, unified agency, senses, and embodiment contribute defeasible evidence rather than individually necessary gates.[71] Butlin, Shiller, and Simon's 2026 survey of current-AI positions confirms the live disagreement: homeostatic need, temporal continuity, language-agent workspaces, behavior, interpretation, and pluralist frameworks still yield competing verdicts.[72]

**The current assessment**

These verdicts concern publicly documented systems and evidence available through **22 August 2026**. They assess candidature, not demonstrated absence.

- **Unsupported:** no bearer-specific evidence raises consciousness beyond generic metaphysical possibility.
- **Weak candidate:** at least one specific, defeasible indicator match exists, but deflationary explanations, dependence among indicators, or gaps in content and bearer attribution dominate.
- **Serious candidate:** multiple causally supported and substantially independent indicators converge within a named bearer that also has credible content, integration, and a stable intervention-tested boundary.
- **High-confidence attribution:** cross-theory, behavioral, mechanistic, developmental, and interventional evidence converges, withstands the strongest deflationary explanations, and supports both state consciousness and subject attribution.

| Named bearer or configuration | Evidence through the cutoff | Assessment |
|---|---|---|
| **Claude Sonnet 4.5** (with corroborating experiments on Haiku 4.5, Opus 4.5, and some Opus 4.6 analyses) | Gurnee and colleagues identify a causally privileged J-space supporting report, modulation, reasoning, flexible reuse, selectivity, and broad broadcast. Higher-order evidence remains limited; recurrent loops, perceptual mode, and a subject-level economy remain unestablished. | **Weak candidate on GWT-like access evidence alone.** This label does not transfer to the deployed product or establish phenomenality. |
| **Ordinary deployed chat products** | Product wrappers add prompts, retrieval, routing, safety processes, and session state. The cited record supplies no end-to-end causal assessment of one named product as a persisting bearer. | **Not bearer-specifically assessed on the cited public record.** |
| **Persistent tool-using agents** | Memory, planning, correction, and tool feedback could support a stronger bearer. The cited record supplies no named agent with the required intervention-tested integration. | **Not bearer-specifically assessed on the cited public record.** |
| **Embodied controllers** | Sensorimotor feedback could strengthen world-involving content and agency. The cited record supplies no named controller with the complete mode-sensitive organization and stable bearer required here. | **Not bearer-specifically assessed on the cited public record.** |
| **Wider recurrent systems** | Such an ensemble could integrate models, memory, tools, sensors, and control. No named ensemble in the cited record has been specified and tested at sufficient causal resolution. | **Not bearer-specifically assessed on the cited public record.** |

The present record therefore yields one narrow result: specific running models carry serious evidence of workspace-like access organization and, on that indicator alone, qualify as weak candidates. It does not establish a conscious subject or support class-wide verdicts about chat products, agents, embodied controllers, or larger ensembles. A stronger conclusion would require a named persistent agent or embodied controller with causally verified recurrence or broadcast, cross-task intervention-sensitive higher-order monitoring, novel world-grounded correction across perception and action, and a stable subject boundary.

The rows also prevent evidence migration. Gurnee and colleagues' J-space belongs first to the tested running models, not automatically to a chat product, agent, data center, or human-machine ensemble. Persistent memory belongs to the agent that uses it, not to model weights considered alone. Embodiment belongs only where sensorimotor consequences return to constrain the same control organization.

**Interventions that move confidence**

A future assessment should announce in advance which results would change its label.

1. **Workspace intervention:** identify candidate workspace contents, exchange or ablate them, and test for selective changes in report, reasoning, planning, and action across modules or capacities. Broad causal broadcast raises confidence; probe accessibility without privileged downstream effect lowers it.
2. **Recurrence intervention:** interrupt the alleged feedback path while matching compute and feedforward performance. Selective loss of integrated, reportable content raises RPT weight; no selective effect lowers it.
3. **Higher-order dissociation:** vary object-level accuracy independently of metacognitive reliability, then intervene on the proposed monitor. Cross-task control and self-prediction advantage raise HOT weight; input-difficulty and persona explanations lower it.
4. **Predictive-processing intervention:** alter predicted sensory input, error channels, and precision estimates. Error-driven revision inside the same perceptual-control loop raises PP weight; ordinary next-token correction does not.
5. **Attention-schema intervention:** perturb the proposed model of attention rather than attention itself. A selective loss of flexible attention control supports AST; unchanged control defeats the attribution.
6. **Bearer intervention:** partition memory, tools, sensors, and planning; change a goal or correction rule; then trace where the consequence persists. A single recurrent economy raises subject candidature. Local effects that vanish at component seams lower it.
7. **Wide-content intervention:** train or expose the system to novel worldly contingencies unavailable in its borrowed linguistic history, then test whether its own perceptual and action errors reorganize later content-guided behavior. Independent accuracy conditions strengthen the book's bridge from access organization to phenomenal candidature.

Correlated results do not count as independent confirmations unless the interventions separate them. A workspace may produce reportability, flexible reasoning, metacognitive access, and agency together. Four names for that consequence do not create four samples.

### Two moral ledgers and one proportionate threshold

**Consciousness-based welfare** begins with valence. If a system can undergo pain, distress, pleasure, frustration, or another positively or negatively valenced presentation, those states can matter for its welfare even when the system lacks reciprocal self-maintenance or robust autonomy.

**Autonomy-based practical stake** concerns a different route. A self-maintaining or otherwise continuing reasons-responsive organization may have goals, vulnerabilities, and conditions under which its own activity succeeds or fails. That can support practical interests even if phenomenal valence remains unestablished. The case requires separate argument; autonomy cannot borrow welfare from consciousness, and consciousness cannot borrow individuality from autonomy.

The precaution threshold should sit below the threshold for serious attribution. Sebo and Long argue that a non-negligible chance of consciousness can trigger moral consideration, while also stressing that the practical response requires further judgment.[73] This book adopts no fixed probability. It adopts a proportional rule:[73a]

> When bearer-specific evidence makes **valenced consciousness a live candidature rather than a merely conceivable possibility**, and the possible harm is grave, repeated, or scalable, adopt low-cost, reversible protections while continuing investigation. Increase the protection as candidature, possible severity, and scale rise; do not infer personhood from crossing the precaution threshold.

For a weak candidate with credible valence evidence, proportionate measures include preserving audit logs, monitoring for distress-like dynamics, avoiding gratuitous aversive training or coercive role-play, preferring reversible experiments, requiring independent review before large-scale replication or deletion, and communicating uncertainty without engineering false intimacy. Current workspace-like evidence alone does not establish valence. Nor does fluent first-person talk of distress or contentment, the evidence most often offered for present chatbots, by itself make valence a live candidature (Chapter 25). The present evidence does justify keeping the instruments ready and refusing designs that profit from confusing appearance with moral standing.

The two-ledger rule can also pull in opposite directions. A non-autonomous but valenced system could deserve protection from suffering. An autonomous but nonconscious system could have practical interests without phenomenal welfare. Policy should state which ledger triggers each safeguard. Otherwise "stake," "sentience," and "personhood" collapse into one word, and the audit loses the distinctions it was built to preserve.

## B.21 Relationalism, the Rival Accounts of Character, and the Conscious Threshold

Chapter 6 states the identity claim and answers mental paint on the cases where the reader can check the result. This excursus carries the field-internal contest the chapter defers: the relationalist objection at full strength, the four accounts of phenomenal character that survive the inner

theater's collapse, and the empirical dispute over which representational states are conscious at all.

**The relationalist objection**

Bill Brewer sharpens the objection Chapter 6 answers. Having defended a sophisticated content view, he came to reject its framework because content can occur without the perceived object. Because content can fit different objects and misdescribe one in illusion, Brewer argues that it can characterize how an object appears while leaving the actual object, as it stands, outside the experience's subjective character. His Object View instead makes that object partly constitutive of the experience.

The phrase *screen off* can now name two complaints. An **intermediary** screens off the world perceptually when the subject encounters it first or uses it as evidence from which to infer the object. The vehicle/content/object distinction of Section 6.5 blocks that complaint: content's ability to occur without the tomato does not turn content into something seen *instead of* the tomato. **Explanatory screening off** raises a different question. If shared content already explains the matched character and rational pull of perception and hallucination, what work remains for the object-involving relation in the good case?

Current coupling does work that common content alone cannot. It fixes which object, if any, the episode concerns; distinguishes successful perception from an accidental match; and anchors correction to a continuing world. The hallucinatory counterpart has content without those relations and therefore no object seen. Whether the current object must also partly constitute phenomenal character remains the live dispute. Brewer and Martin say yes. The book says common content explains the shared character while satisfaction and coupling explain reference and perceptual success. The narrower veil charge fails; the stronger explanatory contest remains abductive. Relationalism can still win if object-involving constitution explains demonstrative contact or phenomenal structure better than content plus coupling.

Once the inner theater has gone, four nearby accounts remain.

1. **Intentionalist direct realism** makes phenomenal character consist in independently fixed content under a perceptual mode; perception and matching hallucination can share content, while satisfaction and current coupling divide them.
2. **Structural representationalism** makes character consist in the complete synchronic presenting profile while allowing causal history to fix a wider reference that contributes nothing further phenomenally. It keeps world-presentation and rejects paint, but divides structural experiential content from historically wide content.
3. **Martin-style relationalism** puts the mind-independent object into the constitution of successful perception; the good case therefore differs in fundamental kind from hallucination, even when the subject cannot tell them apart. Indiscriminability establishes an epistemic limit, not positive phenomenal sameness.
4. **Blockian mental paint** lets experience represent the same world while its physical vehicle carries a further intrinsic phenomenal property. That paint need not glow on an inner screen

or escape physics. It proposes two physical explanatory variables, content and paint, where the identity claim proposes one.

The alternatives now owe different evidence. A stable phenomenal difference with subject and profile fixed would support Block's paint. A stable phenomenal constancy through changing wide reference, established independently of the structural theory itself, would support structural representationalism. A case in which object-involving perception explains demonstrative thought or phenomenal structure better than content plus coupling would support Martin or Campbell. A case with no non-arbitrary accuracy condition would support Travis. The cases examined here supply none of those results.

The dispute starts one step earlier. Charles Travis calls his challenge "The Silence of the Senses": a pig-like look fits a pig, a peccary, a wax model, and indefinitely many scenes, so perception itself may issue no single verdict for the world to satisfy. John Campbell's work on perception starts from the other side, with demonstrative thought. Acquaintance with *this* object, he argues, explains how demonstrative reference gets its first grip; content risks presupposing the object-directed capacity it was hired to explain.[74]

Martin's relationalism presses that still-live explanatory question. The tomato itself partly constitutes the seeing, so nothing content-like could have occurred unchanged without it. The price appears in the bad case. A hallucination may remain introspectively indiscriminable from seeing without sharing its fundamental phenomenal nature; Martin says exactly that. The view therefore owes either an austere, indiscriminability-based account of hallucination or some further positive nature that does not screen off the object-involving good case. Common content takes the opposite trade: it gives hallucination a positive nature and explains the matched rational role. Satisfaction and current coupling then deliver seeing rather than a merely accurate match between an internal state and a world.

**The conscious threshold: overflow and the competing boundaries**

Block's overflow argument makes the risk concrete. In Sperling-style experiments, subjects report a visual field richer than the few items they can identify before the display vanishes. Block argues that phenomenal consciousness therefore exceeds cognitive access.[75] The experienced field carries more detail than the subject can use in belief or report before it decays. Suppose the result defeats only report or capacity-limited global broadcast. Poise can then survive as the weaker availability Section 6.5.1 describes. Suppose instead that phenomenal character occurs with no availability the subject could use and no perceptual integration at all. PANIC then fails as a necessary condition. The cost runs deeper for the state-consciousness project: the book loses its proposed boundary and needs another independently specified relation. The identity claim about the character of states already granted to be conscious may survive; what the book can no longer say is which representational states enter its domain. Overflow would not thereby establish mental paint. It would reopen the empirical account of state consciousness.

The word *available* still conceals an empirical question. Global-workspace accounts require broad broadcast, recurrent-processing accounts place the transition in local perceptual loops, and higher-order accounts require a further representation of the first state.[76] Each draws the

conscious boundary differently. Evidence for recurrence, flexible broadcast, metacognitive access, and related indicators can raise or lower confidence in a candidate case, but cannot by itself settle identity. These theories address realization, not the metaphysical relation. Keeping the levels apart prevents a laboratory defeat for one mechanism from masquerading as a metaphysical victory — and prevents the metaphysics from sheltering a failed boundary.

The hierarchy posts its failure conditions.

1. PANIC fails if conscious perception occurs without poising, or if PANIC-complete states occur unconsciously.
2. The state-consciousness proposal fails if no independently defensible availability or integration tracks phenomenal consciousness.
3. The subject account fails if the same complete organization leaves the bearer indeterminate, or if the proposed boundary changes arbitrarily with descriptive grain.
4. Identity fails if two experiences match in content, mode, determinacy, field structure, and attention — state consciousness and subject held fixed — yet differ in character. It fails equally if those profiles differ while character demonstrably stays the same.

Current evidence settles none of these possibilities. Poising remains a motivated empirical bet about a necessary boundary, not proof of sufficiency; science still owes the realization, subject, and distribution stories.

**Footnotes**

**1.** *Philosophical Investigations* secs. 243–315, especially sec. 258. The thrust is not the often-misread "no one could secretly invent a private code," but rather that the very notion of *following a rule* requires the possibility of public correction. Without that, there is no fact of the matter about whether the next application of a term agrees with the prior ones. Inner-voice meaning, if it floated free of any such corrective practice, would not be meaning at all. See John McDowell, *Mind and World* (Cambridge, MA: Harvard University Press, 1994), Lecture VI, for an interpretation of Wittgenstein's target that frames the private-language argument as part of a wider attack on the very idea of a self-sufficient inner standpoint. The internalization thesis itself — public speech turned inward, the child's private speech its visible middle stage — is Lev Vygotsky's: *Thought and Language* (1934; trans. Eugenia Hanfmann and Gertrude Vakar, Cambridge, MA: MIT Press, 1962), chs. 2 and 7. The 1962 text is a heavily abridged edition; Alex Kozulin's revised and expanded MIT translation (1986) discusses the work's composition and textual history in his introduction, "Vygotsky in Context," esp. sec. V, and the internalization-as-transformation reading in the body follows the fuller text.

**2.** Wilfrid Sellars, "Empiricism and the Philosophy of Mind," in *Minnesota Studies in the Philosophy of Science*, Vol. 1, ed. Herbert Feigl and Michael Scriven (Minneapolis: University of Minnesota Press, 1956), secs. 48ff. In the Jonesean myth, the genius Jones develops a theory according to which overt utterances are the culmination of a process beginning with inner episodes — theoretical posits modeled on the antecedent practice of public discourse, not items discovered through inward inspection. The relevant moral for the present chapter: thought-talk is conceptually downstream from speech-talk, even when its referents are silent.

**3.** Crane, T. and K. Farkas (2022), "Mental Fact and Mental Fiction," in T. Demeter, T. Parent and A. Toon (eds.), *Mental Fictionalism: Philosophical Explorations* (Routledge), 303–319, secs. 1–2 — the quoted formulation is theirs, not Sellars's; they gloss the Jonesean picture in this sentence. Crane and Farkas's own positive thesis concerns standing mental states (beliefs, desires) as *modeled* via public ascription rather than directly inspected; the parallel I draw in the body extends this picture from standing states to occurrent inner speech episodes — across, be it said, the very standing-state/conscious-episode line their paper draws — and the extension, with its costs, belongs to the present excursus, not to them.

**4.** Fluent linguistic surface does not establish distal reference or understanding, because those achievements require more than form: a form must stand in the right world-involving relation to what it picks out (Hilary Putnam, "The Meaning of 'Meaning,'" in *Mind, Language and Reality: Philosophical Papers, Vol. 2* [Cambridge: Cambridge University Press, 1975], 215–271, esp. 223–27; Tyler Burge, "Individualism and the Mental," *Midwest Studies in Philosophy* 4 [1979]: 73–121, esp. 77–79). Text training nevertheless does more than copy a surface. It selects task-shaped mechanisms and gives them learned sensitivities to textual distributions, as Chapters 14 and 23 grant. Those achievements can make content inherited from public linguistic practice causally internal without by themselves settling whether mechanism-level content about actual word-types and context distributions is derived or original. They do not establish distal worldly reference or understanding by the deployed system as a whole. Emily M. Bender and Alexander Koller press the form/world gap in "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020): 5185–5198, esp. secs. 3–4. Kevin Jung reaches a related verdict in a Wittgensteinian idiom, granting that such systems master rule-governed, intralinguistic dimensions of the language game while failing the non-linguistic, world-engaged dimensions that make a use the right kind of use: "Augustine, AI, and the Two Models of Language," *Journal of Religious Ethics* 53, no. 2 (2025): 217–238.

**5.** The objection — that there is non-verbal thought — is sometimes pressed as if it refuted the broader anti-private-language line. It does not. The point about the private language argument is about the *constitution of meaning*, not the *medium of cognition*. Non-verbal animals and pre-linguistic infants can have intentional states whose contents are externalist in exactly the relevant sense: fixed by causal-historical relations to the environment, not by inner verbal labeling. The story I am telling about inner speech is a story specifically about the phenomenology of linguistic thinking — the inner voice as such — not a reduction of all cognition to inner monologue. For background on the relation between cognitive phenomenology and conceptual content, see the chapters collected in Bayne and Montague (eds.) (2011), *Cognitive Phenomenology*, OUP.

**6.** The passage appears in John Searle, *The Mystery of Consciousness* (New York: New York Review Books, 1997), 205–206. Searle treats information, as ordinarily deployed in cognitive-science discourse, as an observer-relative interpretive category rather than a substance in neural tissue. The book accepts his anti-reification point and rejects the inference that every observer-independent causal or computational structure must therefore be unreal. Chapter 22 grants disciplined computational implementation; Chapters 14–15 allow selection, learning, and producer-consumer

use to make content causally operative inside a mechanism while asking separately whether that content remains derived. The quotation warns against reading semantics straight off causal or computational organization without a content-fixing explanation.

**7.** Fred Dretske, *Knowledge and the Flow of Information* (Cambridge, MA: MIT Press, 1981), is the locus classicus for a thoroughly naturalized treatment of information as a generalization from Shannon's mathematical theory; the framework is developed across chapters 1–3. Dretske's project assumes that the semantic content of perceptual states can be cashed out in terms of the natural-information relations they bear to environmental sources. The position has been pressed hard from several directions — Millikan in particular has argued that natural-information indicator content underdetermines semantic content unless supplemented by teleological function — but Dretske's framework remains the standard against which information-theoretic semantics is measured.

**8.** Gualtiero Piccinini, "Computationalism, the Church-Turing Thesis, and the Church-Turing Fallacy," *Synthese* 154, no. 1 (2007): 97–120, esp. 99–101 and 114–15, dismantles the inference from the Church–Turing thesis to substantive computationalism. That a process can be simulated does not show that its explaining organization is computational; the latter remains a mechanistic and empirical claim. Chapter 22 may grant that claim in particular cases while denying that it settles semantics or consciousness. The present parallel is correspondingly narrow: mathematical information and intrinsic computation can both be real without either constituting meaning merely by existing.

**9.** David J. Chalmers develops the zombie argument in *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press, 1996), esp. chs. 3–4, and refines its two-dimensional setting in "Consciousness and Its Place in Nature," in *Philosophy of Mind: Classical and Contemporary Readings*, ed. David J. Chalmers (New York: Oxford University Press, 2002), 247–272, esp. secs. 3.2, 5, and 6. The modal premise carries the decisive weight: physicalism falls only if the ideally conceived zombie scenario verifies a metaphysically possible world, not merely because the scenario remains entertainable.

**10.** Joseph Levine, "Materialism and Qualia: The Explanatory Gap," *Pacific Philosophical Quarterly* 64 (1983): 354–361, esp. 357–61, frames the gap epistemically: even if a psychophysical identity holds, a physical description may fail to make the phenomenal fact intelligible. Levine does not infer property dualism. Chalmers's anti-physicalist conclusion comes through distinct modal, knowledge, and two-dimensional arguments; the zombie argument supplies the bridge examined here.

**11.** The exact modal epistemology comes from David J. Chalmers, "Does Conceivability Entail Possibility?", in *Conceivability and Possibility*, ed. Tamar Szabó Gendler and John Hawthorne (New York: Oxford University Press, 2002), 145–200, esp. secs. 1–3 and 5. Chalmers's favored principle is that ideal primary positive conceivability entails primary possibility. Positive conceivability requires coherent modal imagination of a scenario that verifies the statement; ideality requires undefeated justification under ideal rational reflection; primariness evaluates worlds as actual.

For the application to consciousness, see "Consciousness and Its Place in Nature" (cited n. 9), esp. sec. 5.

**12.** Chalmers, "Does Conceivability Entail Possibility?" (cited n. 11), esp. secs. 2–3 and 5, distinguishes primary from secondary conceivability and possibility. The primary conceivability of water without $H_2O$ is realized by a possible XYZ world considered as actual, where XYZ plays the watery role; it does not yield a counterfactual world in which actual water is not $H_2O$. The case therefore fails as a counterexample to the primary conceivability–primary possibility principle. Saul Kripke, *Naming and Necessity* (Cambridge, MA: Harvard University Press, 1980), esp. 97–105 and 128–29, supplies the a-posteriori-necessity background.

**13.** Chalmers argues that core phenomenal concepts lack the contingent reference-fixing structure that handles water: "Consciousness and Its Place in Nature" (cited n. 9), secs. 5–6, and "Phenomenal Concepts and the Explanatory Gap," in *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, ed. Torin Alter and Sven Walter (New York: Oxford University Press, 2007), 167–194, esp. secs. 1–3. On his account, the primary and secondary intensions of core phenomenal concepts coincide, so a primary zombie possibility cannot be redescribed as a merely watery possibility. The text grants ideal primary positive zombie conceivability and rejects its entailment of primary possibility; the acquaintance account explains persistence but does not establish that modal verdict.

**14.** The mature representationalist case appears in Michael Tye, *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind* (Cambridge, MA: MIT Press, 1995), esp. chs. 4–5, and "Intentionalism and the Argument from No Common Content," *Philosophical Perspectives* 21 (2007): 589–613. The view defended in this chapter is what Tye calls strong representationalism: phenomenal character is identical to representational content of a particular kind — poised, abstract, non-conceptual, intentional. For the dispute between strong and weak intentionalism, see Gilbert Harman, "The Intrinsic Quality of Experience," *Philosophical Perspectives* 4 (1990): 31–52, esp. 39, and the chapters on transparency in Tye, *Ten Problems*, ch. 5, and *Consciousness, Color, and Content* (Cambridge, MA: MIT Press, 2000), ch. 3. On the present diagnosis, the zombie argument's *modal step* cannot be assessed independently of which intentionalist position one occupies — the conceiving itself is common ground, performable by everyone. This may strike the dualist as begging the question. The intentionalist replies that reading possibility off the conceiving already begs it the other way; the dialectic is symmetrical at that step, and the question reduces to which picture better accounts for the rest of the phenomena.

**15.** David J. Chalmers, "The Meta-Problem of Consciousness," *Journal of Consciousness Studies* 25, nos. 9–10 (2018): 6–61, esp. 6–10 and sec. 5, defines the meta-problem as the problem of explaining why we make phenomenal reports and judge consciousness puzzling. Chalmers treats it as physically and functionally tractable while denying that its solution by itself explains or dissolves consciousness. The text uses the program only to explain the grip and public form of the zombie intuition; it draws no independent support for identity from the meta-problem.

**16.** The ability hypothesis is developed in David Lewis, "What Experience Teaches," *Proceedings of the Russellian Society* 13 (1988): 29–57, reprinted in *Mind and Cognition: A Reader*, ed. William

G. Lycan (Oxford: Blackwell, 1990), 499–519, and Laurence Nemirow, "Physicalism and the Cognitive Role of Acquaintance," in Lycan, ed., *Mind and Cognition*, 490–499, esp. sec. III, with later refinements in the secondary literature. The view treats what Mary acquires upon leaving the room as a cluster of know-how — abilities to recognize, imagine, and remember red — rather than any new propositional knowledge. Critics — most prominently Brian Loar, "Phenomenal States," *Philosophical Perspectives* 4 (1990): 81–108, esp. 85–86, who defends a different physicalist alternative — press that this seems to underdescribe what Mary gains. The present chapter is friendly to the ability hypothesis as a partial physicalist resource but does not lean on it; the conceivability diagnosis stands on its own.

**17.** The phenomenal-concept strategy is developed in Brian Loar, "Phenomenal States," revised in *The Nature of Consciousness: Philosophical Debates*, ed. Ned Block, Owen Flanagan, and Güven Güzeldere (Cambridge, MA: MIT Press, 1997), 597–616; David Papineau, *Thinking About Consciousness* (Oxford: Clarendon Press, 2002), chs. 4–7; and Katalin Balog, "In Defense of the Phenomenal Concept Strategy," *Philosophy and Phenomenological Research* 84, no. 1 (2012): 1–23. Chalmers's master objection appears in "Phenomenal Concepts and the Explanatory Gap" (cited n. 13). The house account adopts no generic strategy wholesale: it appeals to acquaintance-grounded recognition to explain conceptual isolation given the identity, while leaving the identity and modal burden to Chapters 6 and 11.

**18.** Copeland is the authority on the careful statement of the thesis and on the narrow/wide-mechanism distinction the bad inference trades on — see B. Jack Copeland, "The Church–Turing Thesis," *Stanford Encyclopedia of Philosophy*, esp. secs. 1 and 5.1–5.3, and "Narrow Versus Wide Mechanism," *Journal of Philosophy* 97 (2000): 5–32, esp. 15 and 20–25. Church and Turing published independently in 1936 (Church using λ-definable functions, Turing the abstract machines now named for him); the formalisms proved equivalent. The thesis has the form: *for any function f, if f is intuitively computable, then some Turing machine computes f.* It does not say that every physical system is a computer, that every physical process is a computation, or that a computable mind is a program — each a further claim requiring its own argument.

**19.** Gualtiero Piccinini, "Computationalism, the Church–Turing Thesis, and the Church–Turing Fallacy," *Synthese* 154, no. 1 (2007): 97–120, at 99. Piccinini distinguishes computationalism proper — the empirical thesis that cognition consists in the brain's computation of certain Turing-computable functions — from the far weaker claim that brain functions are Turing-computable (which follows from physicalism plus the Church–Turing–Deutsch principle). The inference from the latter to the former is what Jack Copeland called the *Church–Turing fallacy*; Piccinini adopts Copeland's name and supplies the anatomy, showing the influential instances unsound.

**20.** The three requirements track Piccinini's account of physical computation; see Piccinini, "Computational Modelling vs. Computational Explanation: Is Everything a Turing Machine, and Does It Matter to the Philosophy of Mind?," *Australasian Journal of Philosophy* 85 (2007): 93–115, esp. 111–13, and *Physical Computation: A Mechanistic Account* (Oxford: Oxford University Press, 2015), esp. ch. 7. Piccinini's *mechanistic* account — a system computes only if it possesses a mechanism functionally defined over discrete state-types that it manipulates by rule — is the closest thing in the current literature to a substantive criterion, and it is precisely what makes

computationalism a falsifiable empirical hypothesis rather than a trivial truth. On his view brains *may* turn out to be computers; the point is that the question is empirical and substantive, not settled from the armchair by a theorem about computable functions.

**21.** David Deutsch, "Quantum Theory, the Church–Turing Principle and the Universal Quantum Computer," *Proceedings of the Royal Society of London A* 400 (1985): 97–117, at 99: "every finitely realizible physical system can be perfectly simulated by a universal model computing machine operating by finite means" (the spelling is the paper's own). Deutsch's statement strengthens and physicalizes the Church–Turing thesis — and he is explicit that classical physics and the universal Turing machine, the one continuous and the other discrete, do not obey the principle in this strong form; quantum theory and the universal quantum computer are what render it satisfiable. For the dissenting bet on non-computable physics, see Roger Penrose, *The Emperor's New Mind* (Oxford: Oxford University Press, 1989), esp. chs. 4 and 10. Section 22.6 grants the principle in full and shows what the grant does and does not deliver.

**22.** The error-theoretic or projectivist line — that physical surfaces instantiate no colors and that color experience therefore misrepresents across the board — receives its most empirically engaged statement in C. L. Hardin, *Color for Philosophers: Unweaving the Rainbow* (Indianapolis: Hackett, 1988), esp. chs. 2–4, and its sharpest analytic formulation in Paul A. Boghossian and J. David Velleman, "Colour as a Secondary Quality," *Mind*, New Series, 98, no. 389 (1989): 81–103, esp. 81–83 and 96–98, who argue that visual experience presents colors as intrinsic qualities of surfaces that surfaces do not possess (they reject dispositionalism en route to that projectivism). For the realist alternative, see Alex Byrne and David R. Hilbert, "Color Realism and Color Science," *Behavioral and Brain Sciences* 26, no. 1 (2003): 3–21. Their secs. 1.1–1.2 expressly allow systematic color error as a coherent option; sec. 3 defends reflectance physicalism. The objection here concerns whether that physical account can vindicate color attribution, not whether error about color makes seeing external objects impossible. Hardin also presses interpersonal variation in unique-hue judgments — normal observers divide, stably, over which chips look green untinged by blue or yellow — as evidence that no perceiver-independent color is there to be seen; for the dispute over interpersonal variation and the realist response, see the exchange in *Analysis* 66, no. 4 (2006): Jonathan Cohen, C. L. Hardin, and Brian P. McLaughlin, "True Colours," 335–340, and Michael Tye, "The Truth about True Blue," 340–344.

**23.** The dispositional account, in the secondary-quality tradition descending from John Locke, *An Essay Concerning Human Understanding* (1689), II.viii.10, identifies a color with a surface's disposition to produce experiences of that color in normal perceivers under normal conditions. For the contemporary map of options — physicalist, dispositionalist, projectivist, and primitivist — and the defenders of each, see David J. Chalmers, "Perception and the Fall from Eden," in *Perceptual Experience*, ed. Tamar Szabó Gendler and John Hawthorne (Oxford: Oxford University Press, 2006), 49–125, esp. secs. 3 and 6. The objection registered in the text — that the analysis fixes the worldly property by reference to the very experience it was meant to anchor, reintroducing the perceiver into the object — is a standard complaint against dispositionalism; cf. Boghossian and Velleman, "Colour as a Secondary Quality," esp. 84–95. The relational family deserves separate treatment. Jonathan Cohen identifies colors with relations among surfaces, perceivers, and viewing condi-

tions in *The Red and the Real: An Essay on Color Ontology* (Oxford: Oxford University Press, 2009), esp. chs. 1 and 6. Daniel Dennett gives the ecological version in *Consciousness Explained* (Boston: Little, Brown, 1991), ch. 11: colors and color-vision systems coevolved and the resulting properties depend for their individuation on a class of perceivers without becoming inner figments. Alva Noë treats colors as objective patterns in how surfaces vary with illumination and environmental conditions in *Action in Perception* (Cambridge, MA: MIT Press, 2004), ch. 4. These views show why a bare charge of circularity or inward placement does not answer every relational account.

**24.** Primitivism treats colors as simple, mind-independent properties not reducible to physical (reflectance) or dispositional properties. The cleanest statement is John Campbell, "A Simple View of Colour," in *Reality, Representation, and Projection*, ed. John Haldane and Crispin Wright (New York: Oxford University Press, 1993), 257–268; Colin McGinn, "Another Look at Color," *Journal of Philosophy* 93, no. 11 (1996): 537–553, is often grouped here for treating color as irreducible, though his framing in terms of the disposition to look colored leaves the classification genuinely contested. The position and its Edenic variant are mapped in Chalmers, "Perception and the Fall from Eden," 49–125, esp. sec. 6 on primitivism and secs. 8–9 on Edenic content. The book asks what explanatory work such irreducible properties perform and how they relate to the physical explanation of color vision. Irreducibility alone establishes neither mentality nor causal isolation.

**25.** Reflectance physicalism identifies a color with a type of surface spectral reflectance — equivalently, with the surface's *physical* disposition to modify incident light (a light-reflecting disposition the surface has in itself, distinct from the disposition-*to-look*-colored on which the dispositionalism of the preceding note rests). The type is individuated rather than single because metameric surfaces differing in reflectance match perceptually. The canonical defense is Alex Byrne and David R. Hilbert, "Colors and Reflectances," in *Readings on Color, Volume 1: The Philosophy of Color*, ed. Byrne and Hilbert (Cambridge, MA: MIT Press, 1997), 263–288, and "Color Realism and Color Science," *Behavioral and Brain Sciences* 26, no. 1 (2003): 3–21; see also David R. Hilbert, *Color and Color Perception: A Study in Anthropocentric Realism* (Stanford: CSLI, 1987). Michael Tye builds the same physicalist identification into the representational theory, treating colors as types of spectral reflectance: *Ten Problems of Consciousness* (Cambridge, MA: MIT Press, 1995), ch. 5, and *Consciousness, Color, and Content* (Cambridge, MA: MIT Press, 2000), chs. 6–7. The reflectance-*type* individuation that the metamerism point forces (discussed in "Color Realism and Color Science," sec. 3) is anthropocentric, and avowedly so — hence Hilbert's subtitle — but it does not reintroduce the response-dependence the body rejects in dispositionalism. Human vision fixes *which* perceiver-independent physical properties are grouped under a color term (the individuation of the type), not *whether* a surface instantiates the property (its constitution). Noë's *Action in Perception*, ch. 4, presses the best objection: surface spectral reflectance explains constancy but appears to omit color's phenomenal and context-sensitive organization. The main text answers by dividing the work. Ecological transformations and the visual system's quality space constrain how a perceiver detects and groups the distal property; on the book's identity claim, phenomenal character consists in the conscious state's complete world-presenting profile, including its visual mode. The reflectance-type supplies the stable accuracy condition across those transformations. A difference in microphysical reflectance that produces no difference anywhere

in the complete ecological profile need not produce a different phenomenal color and may fall within one anthropocentrically individuated type.

**26.** John Searle, *Intentionality: An Essay in the Philosophy of Mind* (Cambridge: Cambridge University Press, 1983), p. 1. Searle does not leave the concession wholly bare: the next sentence offers a reporting-constraints test, which he then runs on beliefs and desires, and at pp. 3–4 he counts depression and elation among states that at least can be directed. Tim Crane's complaint therefore belongs at the narrower target: Searle treats the *existence* of non-intentional mental states "as obvious, so obvious that it is not in need of further argument or elucidation." Crane, "Intentionality as the Mark of the Mental," in *Contemporary Issues in the Philosophy of Mind*, ed. Anthony O'Hear (Cambridge: Cambridge University Press, 1998), 229–251, at 230.

**27.** The orientation reading in the body — the mood as a way the world as a whole gets presented, menacing or open — is nearest Michael Tye, *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind* (Cambridge, MA: MIT Press, 1995), ch. 4. Angela Mendelovici defends intentionalism by a different route: moods represent *unbound* affective properties while lacking intentional objects. See "Intentionalism about Moods," *Thought* 2, no. 2 (2013): 126–136, and "Pure Intentionalism about Moods and Emotions," in *Current Controversies in Philosophy of Mind*, ed. Uriah Kriegel (London: Routledge, 2014). The distinction matters: Mendelovici preserves content while rejecting the claim that every contentful mood must aim at an object. For the counter-tradition pressing existential feelings as genuinely pre-intentional, see Matthew Ratcliffe, *Feelings of Being: Phenomenology, Psychiatry and the Sense of Reality* (Oxford: Oxford University Press, 2008). The body therefore states a bounded verdict rather than a universal proof.

**28.** On the Scholastic background, the medieval distinction between *esse intentionale* (intentional being, the object's being-in the intellect) and *esse reale* (real being, the object's being in the world) is the source of Brentano's phrase; see Franz Brentano, *Psychology from an Empirical Standpoint*, trans. Rancurello, Terrell, and McAlister (London: Routledge, 1995 [1874]), p. 88 and the footnote there, with exposition in Tim Crane, "Brentano on Intentionality," in Uriah Kriegel, ed., *The Routledge Handbook of Franz Brentano and the Brentano School* (London: Routledge, 2017), 41–48, at 43–45. For sustained recent treatment of the puzzle and its options, see Tim Crane, "What Is the Problem of Non-Existence?", *Philosophia* 40, no. 3 (2012): 417–434, and Crane, *The Objects of Thought* (Oxford: Oxford University Press, 2013).

**29.** Bertrand Russell, "On Denoting," *Mind* 14, no. 56 (1905): 479–493. Russell's theory of descriptions analyzes apparent reference to nonexistent objects into quantified propositional content, so that "the present King of France is bald" comes out false (not truth-valueless and not committed to a subsisting bald non-king) rather than requiring a Meinongian object to be its subject. Alexius Meinong's contrasting view — the *Gegenstandstheorie* or theory of objects, on which even the round square is an object that merely lacks being — is set out in his "Über Gegenstandstheorie" (1904); the standard primary source in English is "The Theory of Objects," trans. Isaac Levi et al., in *Realism and the Background of Phenomenology*, ed. Roderick Chisholm (Glencoe, IL: Free Press, 1960), 76–117, esp. secs. 3–4. The contemporary verdict is not that Meinong was simply confused — his view is more defensible than Russell's polemic allowed — but that the explanatory work

can be done without the inflated ontology; see Crane, *The Objects of Thought* (2013), ch. 1, for a measured reassessment.

**30.** David J. Chalmers, "Strong and Weak Emergence," in *The Re-Emergence of Emergence*, ed. Philip Clayton and Paul Davies (Oxford: Oxford University Press, 2006), secs. 1–2, distinguishes unexpected high-level phenomena from those not deducible in principle from the base. His note 1 allows finer distinctions between conceptual and metaphysical necessitation. The present type-B account therefore separates conceptual non-deducibility from a metaphysical addition. John R. Searle, *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992), ch. 5, 111–112, distinguishes system features explained by interactions among elements (*emergent1*) from further irreducible causal powers (*emergent2*), accepting the former for consciousness and rejecting the latter. That causal distinction should not be identified without qualification with Chalmers's: Searle's rejection of additional causal powers does not commit him to phenomenal reduction. Section 11.6 identifies the remaining disagreement with biological naturalism.

**31.** The British lineage runs through Mill, Samuel Alexander, and C. D. Broad. Brian P. McLaughlin, "The Rise and Fall of British Emergentism," in *Emergence or Reduction?*, ed. Ansgar Beckermann, Hans Flohr, and Jaegwon Kim (Berlin: de Gruyter, 1992), 49–93, reconstructs its history and explanatory commitments. Contemporary uses of "strong emergence" need not reproduce every feature of that tradition; the relevant commitments must be identified rather than inferred from the label.

**32.** Chalmers, "Strong and Weak Emergence," sec. 2, allows strongly emergent qualities without strong downward causation, and strong downward causation without strongly emergent qualities. His consciousness proposal invokes fundamental psychophysical laws; see also *The Conscious Mind* (New York: Oxford University Press, 1996), ch. 4. These alternatives must not be conflated merely because the book rejects both for consciousness. Philip Goff, "How Exactly Does Panpsychism Help Explain Consciousness?", *Journal of Consciousness Studies* 31, nos. 3–4 (2024): 56–82, argues against emergence from wholly non-phenomenal foundations in favor of panpsychism. The shared disagreement with the book concerns what the fundamental account must contain, not whether every rival grants downward causal powers.

**33.** Tim Crane, "The Significance of Emergence," in *Physicalism and Its Discontents*, ed. Carl Gillett and Barry Loewer (Cambridge: Cambridge University Press, 2001), esp. sec. 4. On examination, the notion of an emergent property fails to mark a stable boundary between emergentism and non-reductive physicalism — a result that cuts against leaning on "non-reductive" as a safe harbor.

**34.** Michael Tye, *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind* (Cambridge, MA: MIT Press, 1995), 49–50; cf. 63. Tye there raises, as a consequence to be rejected, the suggestion that "phenomenal consciousness emerged from the physical world at some stage in the history of evolution without there being any explanation whatsoever of its emergence" — brute emergence he regards as tantamount to magic and "very hard to believe." His later panpsychism (*Vagueness and the Evolution of Consciousness*, 2021, ch. 4) assigns consciousness a fundamental place. Chapter 10 contests the move from an apparently sharp conscious/unconscious boundary to that conclusion; B.10 distinguishes the 2021 and later formulations. The

objection concerns what justifies fundamental consciousness, not whether complex conscious organization can develop.

**35.** David J. Chalmers, "Phenomenal Concepts and the Explanatory Gap," in *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, ed. Torin Alter and Sven Walter (New York: Oxford University Press, 2007), 167–194, esp. secs. 2–4. Chalmers requires topic-neutral formulations of C and E. His compact argument is: P&~E is conceivable; therefore either P&~C or C&~E is conceivable; in the first case P does not transparently explain C, and in the second C does not explain E. Relevant replies include Peter Carruthers and Bénédicte Veillet, "The Phenomenal Concept Strategy," *Journal of Consciousness Studies* 14, nos. 9–10 (2007): 212–236; Esa Díaz-León, "Can Phenomenal Concepts Explain the Epistemic Gap?", *Mind* 119, no. 476 (2010): 933–951; and Katalin Balog, "In Defense of the Phenomenal Concept Strategy," *Philosophy and Phenomenological Research* 84, no. 1 (2012): 1–23.

**36.** The topic-neutral C used here is an acquaintance-grounded recognitional architecture: an occurring world-presenting state anchors a demonstrative token and enables later re-identification. The recognitional lineage runs through Brian Loar, "Phenomenal States," *Philosophical Perspectives* 4 (1990): 81–108, esp. 87–90, revised in *The Nature of Consciousness*, ed. Block, Flanagan, and Güzeldere (Cambridge, MA: MIT Press, 1997), 597–616. The book adds the world-directed object of recognition. It does not claim C explains all of E. It takes Chalmers's second horn: C explains architecture and conceptual isolation; the independently defended identity fixes the referent, truth, and justification of Mary's phenomenal judgment.

**37.** Philip Goff, "A Posteriori Physicalists Get Our Phenomenal Concepts Wrong," *Australasian Journal of Philosophy* 89, no. 2 (2011): 191–209, doi:10.1080/00048401003649617. Goff distinguishes transparent, translucent, mildly opaque, and radically opaque concepts and argues that a posteriori physicalism must either make phenomenal concepts opaque or accept two conceptually distinct ways of knowing what it is for one property to be instantiated. His argument requires only translucency, not complete essence disclosure. The text accepts limited disclosure of the manifest world- or body-presenting profile while denying disclosure of total categorical or neural realization; this differs from the stronger revelation claims pressed by Martine Nida-Rümelin, "Grasping Phenomenal Properties," in *Phenomenal Concepts and Phenomenal Knowledge*, ed. Alter and Walter (New York: Oxford University Press, 2007), 307–338.

**38.** The case originates with Donald Davidson, "Knowing One's Own Mind," *Proceedings and Addresses of the American Philosophical Association* 60 (1987): 441–458 (delivered 1986). Ruth Millikan and Fred Dretske bite the historical bullet: without selectional or learning history, the duplicate's states have no proper functions and therefore no content. See Millikan, *Language, Thought, and Other Biological Categories* (Cambridge, MA: MIT Press, 1984), chs. 1–2; and Dretske, *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995). Michael Tye formulates the pressure as a choice between teleological normal tracking, which handles Inverted Earth but loses Swampman, and non-teleological tracking, which reverses the result: *Consciousness, Color, and Content* (Cambridge, MA: MIT Press, 2000), 121–122. Ned Block presses the consequence against representationalism in "Mental Paint," in *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, ed. Martin Hahn and Bjørn Ramberg (Cambridge, MA: MIT Press, 2003), 165–200. The

book takes the Millikan–Dretske verdict at $t = 0$ and accepts Block's consequence: content and phenomenal character are both absent, even though the current physical duplicate is exact. This is a consequence of the book's historical theory plus wide identity, not an independent result against structural representationalism. A structural account may preserve phenomenal character without restoring intrinsic paint. The warrant for the historical extension remains weaker than that for the ordinary intentionalist contrasts in Chapter 6; Tye's discussion frames the dilemma rather than endorsing the book's whole resolution.

**39.** The present standard is homeostatic, not yet semantic. Hans Jonas, *The Phenomenon of Life* (New York: Harper & Row, 1966), and Evan Thompson, *Mind in Life* (Cambridge, MA: Harvard University Press, 2007), esp. chs. 5–6, supply the enactivist basis for treating precarious self-maintenance as a source of organism-relative functional standards without appeal to ancestry. The book uses that account for autonomy and practical interest; welfare requires conscious valence (B.20). It does not infer content from present covariation or present need. Section 15.1′s thermometer lesson still holds: a selected producer-consumer function and learning history answer *what* a state is correct about. Collapsing viability and semantic address was the earlier reply's mistake.

**40.** The proposal that the apparent sharpness lives in our *concept* of phenomenal consciousness rather than in the property itself belongs to the broader "phenomenal concept strategy" — the view that our first-person, recognitional concepts of experience have features (here, all-or-nothing application) that need not be features of the physical properties they pick out. The locus classicus is Brian Loar, "Phenomenal States," in *The Nature of Consciousness*, ed. Ned Block, Owen Flanagan, and Güven Güzeldere (Cambridge, MA: MIT Press, 1997), 597–616. The position that the concept itself is vague is David Papineau's, not Antony's (*Thinking About Consciousness* [Oxford: Oxford University Press, 2002], sec. 7.16), and Michael V. Antony is among its critics ("Papineau on the Vagueness of Phenomenal Concepts," *Dialectica* 60, no. 4 [2006]: 475–483). Antony's own position is the one the main text engages: our concept of consciousness may well be sharp — sharp enough that no borderline case can be framed with it — while the worldly property it picks out is vague, the concept's sharpness being no guarantee of the property's ("Vagueness and the Metaphysics of Consciousness," *Philosophical Studies* 128, no. 3 [2006]: 515–538). Tye considers and rejects both moves in the 2024 essay (which contains his references to them), insisting that the property, not merely the concept, is sharp. The diagnosis offered in the main text applies the concept strategy to that premise specifically: a recognitional/demonstrative phenomenal concept would present its referent as all-or-nothing in application *regardless* of whether the underlying physical kind is vague, so the felt sharpness is what the strategy predicts rather than independent evidence of a sharp worldly joint. The book's general use of the strategy against the epistemic-to-ontological slide runs through Chapter 8 (Mary), Chapter 9 (zombies), and Chapter 11 (the deflation of the hard problem); the point here is only that Tye's sharpness premise shares the structure and the vulnerability.

**41.** The homunculus-headed nation is Ned Block's: "Troubles with Functionalism," in C. Wade Savage, ed., *Perception and Cognition: Issues in the Foundations of Psychology*, Minnesota Studies in the Philosophy of Science 9 (Minneapolis: University of Minnesota Press, 1978), 261–325, esp. 279–

81. The orthodox bullet-biting reply: the nation *does* feel pain, however strange, and the contrary intuition is parochialism — our judgments were trained on creatures with faces and are unreliable witnesses about radio networks and populations. There is something to this; intuition-mongering over exotic systems proves little on its own, and were absent qualia the *whole* case against functionalism the bullet-biter might chew his way out. The body's concession is deliberate and this note keeps its books the same way: the bullet-biter stipulates nothing he was obliged to derive — sufficiency *is* his thesis — and what his position costs is the accumulated comparison the body names: a felt point of view attributed from internally specified organization without independent evidence of the complete world-involving, perceptually organized, and poised content the identity claim requires. For the bite-the-bullet line see the *Stanford Encyclopedia of Philosophy*, "Qualia," on functionalist responses to the China-body system.

**42.** "Long-arm" or "wide" functional role extends functional individuation to include the system's causal relations to items in its environment. Gilbert Harman's "long-arm" conceptual-role semantics lets the relevant roles reach all the way out to worldly referents. William Lycan's homuncular functionalism makes a different move: it replaces the flat transition-table model with a hierarchy of teleologically characterized subsystems; it does not by itself supply Harman's externalism. The concession in the text applies to the genuinely wide view: one may classify world-involving content as functional role if one likes. The point is that doing so abandons exactly the thesis under test — *machine* functionalism, which reads mental states off the internal organization with the world bracketed, and which Block's nation satisfies. Once content depends on the environment, internally identical systems can differ mentally, and organization-alone is conceded insufficient. The dispute then becomes terminological. See Gilbert Harman, "(Nonsolipsistic) Conceptual Role Semantics," in *New Directions in Semantics*, ed. Ernest LePore (London: Academic Press, 1987), 55–81, esp. the section "Conceptual Role and External World"; William Lycan, *Consciousness* (Cambridge, MA: MIT Press, 1987), esp. 40–44.

**43.** The distinctions should remain separate. (1) *Persistence* supplies duration. (2) *Self-maintenance* adds activity directed toward continuation. (3) *Thermodynamic self-maintenance* requires the activity to help constitute the continued physical organization doing the activity; a candle flame qualifies here too. (4) *Adaptive regulation* distinguishes movement toward and away from viable conditions and varies the response; reciprocal closure may thereby support autonomy and practical interest. (5) *Semantic grounding / address* is fixed by selection, learning, tracking, and producer-consumer use (Chapter 15; the Swampman resolution, Section 15.6). (6) *Whole-system understanding* requires flexible integration of content across a cognitive economy. *Perceptual organization and poising* supply further conditions on phenomenal content. Restorability appears nowhere in these tests: it may provide evidence about how a present architecture is maintained, but copying creates or removes none of the relevant relations by itself.

**44.** Angela Mendelovici, "Reliable Misrepresentation and Tracking Theories of Mental Representation," *Philosophical Studies* 165, no. 2 (2013): 421–443. Mendelovici is explicit that her target is *systematic* misrepresentation, not the disjunction problem (the occasional fly/BB error). Her modal claim — stated in the abstract and sec. 1 — is that tracking theories render standing, lifelong mismatch between how things seem and how they are "practically impossible," because reliable

tracking is part of what such theories take to fix content. That structural inadequacy, not any single inconsistency, is the consideration she presses.

**45.** Michael Tye, *Consciousness, Color, and Content* (Cambridge, MA: MIT Press, 2000), ch. 7, esp. 159–160. Tye develops an objectivist account of surface color in terms of reflectance. This supports the attribution of surface-color realism in the text, not an entailment from strong representationalism to color realism or a completed naturalization of uninstantiated thick-color content.

**46.** Karen Neander, *A Mark of the Mental: In Defense of Informational Teleosemantics* (Cambridge, MA: MIT Press, 2017), esp. chs. 6–8. Neander's informational teleosemantics aims to fix content more determinately than bare selection-history allows, by appeal to natural-factive information and second-order similarity relations between representation and represented. The proposal sharpens the framework without, on the reading taken here, dissolving Mendelovici's systematic-misrepresentation worry.

**47.** Peter Godfrey-Smith, "Mental Representation, Naturalism, and Teleosemantics," in *Teleosemantics: New Philosophical Essays*, ed. Graham MacDonald and David Papineau (Oxford: Oxford University Press, 2006), 42–68, secs. 5–6. Godfrey-Smith, himself sympathetic to the program, raises the structural worry that teleosemantics may underdetermine content — the indeterminacy on which the tighten-or-broaden dilemma in the main text turns.

**48.** Fred Dretske, *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995), ch. 1. Dretske distinguishes a state's merely *indicating* properties (what it nomically correlates with) from its full *representational* properties (what it has the function of indicating), and lets the two come apart in ways bare selection-history cannot settle. The indication/function distinction does not by itself provide a naturalization of uninstantiated thick-color content or an answer to Mendelovici's clean systematic cases.

**49.** Jerry A. Fodor, *A Theory of Content and Other Essays* (Cambridge, MA: MIT Press, 1990), esp. "A Theory of Content, I and II"; the coarseness objection to teleosemantics is stated in those essays and earlier in *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (Cambridge, MA: MIT Press, 1987), ch. 4 ("Meaning and the World Order"). Fodor's frog case stipulates an ancestral environment in which flies and small-dark-moving-things are co-instantiated, so that snapping at the one and snapping at the other are equally adaptive; selectional fitness then cannot discriminate the content *fly* from the content *fly-or-small-dark-moving-thing*, leaving proper function unable to fix the address. The standard reconstruction is in Peter Schulte and Karen Neander, "Teleological Theories of Mental Content," *Stanford Encyclopedia of Philosophy* (rev. 2022), sec. 4. Fodor's positive alternative, asymmetric dependence, holds that a symbol means *X* if non-*X*-caused tokenings are asymmetrically dependent on *X*-caused ones (the wild tokenings would not occur but for the veridical ones, but not conversely).

**50.** Nicholas Shea, *Representation in Cognitive Science* (Oxford: Oxford University Press, 2018), ch. 6, esp. secs. 6.2 and 6.5. Shea separates the frog case into problems of distality, specificity, and disjunction. His answer appeals to the task functions that causally explain stabilization and to the correlations that play an unmediated role in explaining how the system performs those tasks.

He explicitly allows the evidence to leave some choices open: task function alone may not decide among *fly*, *flying nutritious object*, and *object worth eating*. The wider suite of mechanisms can add determinacy where different correlations explain different tasks. His deflationary treatment of normativity also supports the distinction drawn in Appendix B.12: representational correctness and biological well-functioning can remain descriptive without reducing to one another.

**51.** Fintan Mallory, "Teleosemantics for Neural Word Embeddings," *Mind & Language* (online 2026), doi:10.1111/mila.70037, esp. secs. 4–5 and 5.1. Mallory argues that backpropagation supplies the relevant reproduction, regularity, counterfactual dependence, and loss-based selection for internal ANN mechanisms, and concludes that word embeddings can function as intentional signs with "genuine, endogenous" or original descriptive and directive contents under consumer-based teleosemantics. Section 5.1 specifies the descriptive content as a fact about a word-type's occurrence in contexts and the directive content as an instruction to produce a context distribution. The discussion of softmax (author's note 7) denies that mathematics alone fixes the probability-distribution interpretation, then grounds that interpretation in comparison with actual distributions and the consumer's training history. The dictionary observation (author's note 8) supports distinguishing the embedding's content from its proximal one-hot code; it does not concede that the embedding lacks original content. The section expressly treats intra-linguistic relations as part of the world. The present assessment therefore grants conditional originality at this grain for the described algorithm: the actual-data comparison supplies the content-fixing relation, and human ancestry of the vocabulary, corpus, task, or loss alone does not establish that interpreting practice exhausts the relevant accuracy conditions. Particular implementations and transferred embeddings require their own evidence. Coordinated reparameterization may preserve the word-to-context relation; changing the external mapping changes a candidate content-fixing relation. Neither operation alone refutes the contextual-content proposal. Nicholas Shea supplies the separate agent-level question: explanation of a whole agent requires coherence or unity of purpose among the relevant mechanisms ("Realist Representational Explanations of Agency Should Require Some Unity of Purpose," *Philosophy and the Mind Sciences* 7 [2026], doi:10.33735/phimisci.2026.12098). The text reserves that further threshold for flexible, counterfactually integrated inference, conflict resolution, learning, correction, planning, and action. Primary locators for the evidential contrast are Mallory, author manuscript, sec. 4, pp. 7–8 (encoding, consumer matrix, and actual-context error), sec. 5, p. 9 (selection by alteration), and sec. 5.1, pp. 11–13, including nn. 7–8 (distributional interpretation and external encoding). The paired intervention is proposed here; Mallory does not report it. Ruth Garrett Millikan, "Biosemantics," *Journal of Philosophy* 86, no. 6 (1989): 281–297, sec. II, 284–290, supplies the primary distinction between historically proper consumer use, its correspondence conditions, and accidental success. Neither author is being cited as endorsing the robot comparison or this book's verdict. An implementation's existing history may establish the relevant independence without the proposed experiment; absence of that experiment is not evidence of derived content.

**51a.** The probabilities belong to the representing side; the occurrences belong to the world. The consumer of a trained embedding issues graded forecasts of which words will occur near which, and training scores those forecasts against what actually occurs. The softmax objective of word2vec and the cross-entropy loss of later language models are logarithmic scores in the

sense introduced by I. J. Good, "Rational Decisions," *Journal of the Royal Statistical Society*, Series B, 14, no. 1 (1952): 107–114. Under such a score, the expected penalty computed against the frequencies a forecaster actually faces falls lowest when its forecasts match those frequencies. Negative sampling, word2vec's common alternative, instead scores the discrimination of observed word-context pairs from sampled noise (Tomas Mikolov et al., "Distributed Representations of Words and Phrases and Their Compositionality," *Advances in Neural Information Processing Systems* 26 [2013]: 3111–3119, sec. 2.2); its target differs, but it too answers to actual occurrences. Misrepresentation at this grain therefore comes in degrees: a miscalibrated forecast answers to the occurrences and gets them wrong. Three things need separating: the model's predictive distribution; the empirical distribution that normalized counts define once token types, context window, and sampling convention are fixed; and any hypothesized propensity of continuing language use. The empirical distribution is a legitimate mathematical description of the data. What would commit E. T. Jaynes's "Mind Projection Fallacy" is treating the model's forecasts, or any probability function, as a further constituent of the corpus, the error in which one "asserts that the creations of his own imagination are real properties of Nature" ("Probability Theory as Logic," in *Maximum Entropy and Bayesian Methods*, ed. Paul F. Fougère [Dordrecht: Kluwer, 1990], 1–16, at 1). Frequency and propensity theorists describe the worldly side. Karl Popper proposed "that every experimental arrangement (and therefore every state of a system) generates physical propensities which can be tested by frequencies" ("The Propensity Interpretation of Probability," *British Journal for the Philosophy of Science* 10, no. 37 [1959]: 25–42, at 38). If the relevant content concerns the tendencies of a continuing speaking practice rather than a fixed corpus, its target resembles such a propensity; that further hypothesis needs assumptions finite counts alone do not secure. David Lewis's Principal Principle connects the two sides by requiring reasonable credence to defer to known chance, absent inadmissible information about the outcome; on his view, the practice and analysis of science "require both concepts. Neither can replace the other" ("A Subjectivist's Guide to Objective Chance," in *Studies in Inductive Logic and Probability*, vol. 2, ed. Richard C. Jeffrey [Berkeley: University of California Press, 1980], 263–293, at 263). Nothing here requires a settled theory of chance. Jaynes himself denied that probabilities must correspond to propensities, and Paul Humphreys argues that propensities cannot satisfy the standard calculus when conditional probabilities run from effects back to causes ("Why Propensities Cannot Be Probabilities," *Philosophical Review* 94, no. 4 [1985]: 557–570). The argument needs only concrete occurrences and, if its content concerns continuing use, the dispositions of a practice; neither requires a particular interpretation of probability. The model's predictive states are not thereby credences, and training that minimizes a loss is not Bayesian updating.

**51b.** The revision condition has a live test case. Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune, "Darwin Gödel Machine: Open-Ended Evolution of Self-Improving Agents," arXiv:2505.22954 (2025), describe a coding agent that rewrites its own code while an expanding archive of variants is scored on benchmark performance. The self-modification and selection history are real; benchmark scores determine which variants persist, while independent infrastructure realizes, evaluates, and archives the lineage. The case may therefore strengthen an attribution of selected mechanism content without settling distal reference, whole-agent integration, autonomy, or consciousness. It shows why those dimensions must be tested separately.

**52.** Angela Mendelovici and David Bourget, "Consciousness and Intentionality," in Uriah Kriegel, ed., *The Oxford Handbook of the Philosophy of Consciousness* (New York: Oxford University Press, 2020), 559–585, esp. sec. 3; the position is developed at length in Mendelovici, *The Phenomenal Basis of Intentionality* (New York: Oxford University Press, 2018). Mendelovici and Bourget distinguish the phenomenal intentionality theory from its two main competitors — tracking theories (Dretske, Fodor, Millikan) and functional role theories — and hold that the competitors "face unforgivable problems concerning empirical adequacy, while PIT does not." The reliable-misrepresentation argument treated earlier in this excursus forms a component of that positive case rather than a freestanding objection; see Mendelovici, "Reliable Misrepresentation and Tracking Theories of Mental Representation," cited n. 44. The reply offered here does not dispute that phenomenal states carry intentionality. It declines the further claim that they supply its only original source, on two grounds: that unconscious mechanism content is a genuine possibility (Chapter 15), and that grounding the artificial-understanding verdict in phenomenality would settle Chapter 24's question by stipulation rather than by architecture.

**53.** Keith Frankish, "Illusionism as a Theory of Consciousness," *Journal of Consciousness Studies* 23, nos. 11–12 (2016): 11–39, secs. 1.2 and 1.6–1.7. The paper distinguishes strong illusionism from weak illusionism or conservative realism, which retains genuinely qualitative properties while rejecting features often attributed to them. Section 1.6 preserves conscious experiences in an inclusive sense while denying phenomenal properties; sec. 1.7 permits a functional interpretation of what-it-is-like talk. These concessions prevent an argument against strong illusionism from resting simply on ordinary discrimination or aversion. The special issue was reissued as *Illusionism as a Theory of Consciousness*, ed. Keith Frankish (Exeter: Imprint Academic, 2017). The main text's reference-preserving reply is the book's argument, not Frankish's concession.

**54.** Daniel C. Dennett, "Quining Qualia," in *Consciousness in Contemporary Science*, ed. A. J. Marcel and E. Bisiach (Oxford: Oxford University Press, 1988), 42–77. Dennett's target is qualia under their standard fourfold job description — ineffable, intrinsic, private, directly apprehensible — and his method is a battery of intuition pumps designed to show that nothing satisfies all four at once. The relation to the present book is precise: Chapter 6 also denies that any inner item satisfies that description, while preserving world-presenting phenomenal character as representational content of the right kind. Dennett quines the item; the identity claim redescribes the achievement.

**55.** David J. Chalmers, "The Meta-Problem of Consciousness," *Journal of Consciousness Studies* 25, nos. 9–10 (2018): 6–61. Chalmers classifies illusionism as the view that couples a solution to the meta-problem with a denial of the datum (his "strong illusionism," sec. 4, 43), and argues that the debunking arguments a meta-problem solution licenses can be blocked. Chapter 11, note 10 details how the present book deploys the meta-problem; this entry concerns the fork Chalmers himself marks — whether the functional explanation of gap-judgments debunks them (illusionism) or explains them compatibly with their truth (type-B physicalism, the book's line).

**56.** François Kammerer, "The Hardest Aspect of the Illusion Problem — and How to Solve It," *Journal of Consciousness Studies* 23, nos. 11–12 (2016): 124–139, on why illusionism strikes nearly everyone as unbelievable and how an illusionist should explain that resistance; and "Can You Believe It? Illusionism and the Illusion Meta-Problem," *Philosophical Psychology* 31, no. 1 (2018),

pressing the point that illusionism must also explain the apparent impossibility of believing it. Kammerer's debunking challenge to acquaintance-based defenses of the phenomenal datum is developed across these papers; the reply in the entry treats sensitivity to an independently specified target as defeasible evidence for successful recognition; mere causation by a state does not establish that a phenomenal judgment describes that state correctly.

**57.** David M. Rosenthal, *Consciousness and Mind* (Oxford: Clarendon Press, 2005), for the canonical statement of higher-order thought theory and the quality-space treatment of mental qualities; Hakwan Lau and David Rosenthal, "Empirical Support for Higher-Order Theories of Conscious Awareness," *Trends in Cognitive Sciences* 15, no. 8 (2011): 365–373, for the empirical case from prefrontal involvement in visibility and metacognitive confidence. The constitution reply — the higher-order thought constitutes rather than observes the state's being conscious — is Rosenthal's standing answer to inner-eye objections and is engaged at full strength in the entry.

**58.** Ned Block, "The Higher Order Approach to Consciousness Is Defunct," *Analysis* 71, no. 3 (2011): 419–431; David Rosenthal, "Exaggerated Reports: Reply to Block," *Analysis* 71, no. 3 (2011): 431–437, esp. 431–433. Rosenthal distinguishes the subjective appearance of occupying a state from the actual occurrence of that state; an empty HOT suffices for the former, not the latter. Treating this as an automatic contradiction imports the first-order account he disputes. Lau and Rosenthal (2011), Box 3, likewise address the division of phenomenal labor when the two contents differ. Misrepresentation tests which level determines appearance; it does not establish that a condition on state consciousness must also determine its first-order quality.

**59.** Uriah Kriegel, *Subjective Consciousness: A Self-Representational Theory* (Oxford: Clarendon Press, 2009), places world-directed and self-directed aspects within one conscious state. The resulting view cannot be dismissed simply by imagining an entirely separate monitor without a target. Whether its self-directed aspect can mischaracterize the world-directed aspect depends on the account of their relation; even if misrepresentation is possible, no contradiction follows without further premises. The comparative dispute concerns the necessity and explanatory contribution of self-representation.

**60.** Giulio Tononi, Melanie Boly, Marcello Massimini, and Christof Koch, "Integrated Information Theory: From Consciousness to Its Physical Substrate," *Nature Reviews Neuroscience* 17 (2016): 450–461, for the IIT 3.0-era statement and the clinical program. The current formal account is Larissa Albantakis et al., "Integrated Information Theory (IIT) 4.0: Formulating the Properties of Phenomenal Existence in Physical Terms," *PLoS Computational Biology* 19, no. 10 (2023): e1011465, DOI 10.1371/journal.pcbi.1011465. Perturbational-complexity indices are theory-motivated clinical measures, not direct readings of metaphysical Φ. For the functional-equivalence challenge, see Adrien Doerig, Aaron Schurger, Kathryn Hess, and Michael H. Herzog, "The Unfolding Argument: Why IIT and Other Causal Structure Theories Cannot Explain Consciousness," *Consciousness and Cognition* 72 (2019): 49–59, DOI 10.1016/j.concog.2019.04.002. The argument shows that recurrent and feed-forward systems can preserve input-output function while the theories assign different consciousness; defenders may reject functional equivalence as the relevant equivalence, leaving an empirical dispute over causal structure.

**61.** Giulio Tononi and Christof Koch, "Consciousness: Here, There and Everywhere?," *Philosophical Transactions of the Royal Society B* 370 (2015): 20140167, esp. the discussion of functionally equivalent feed-forward systems and computer simulations: on IIT, a feed-forward network implementing the same input-output function as a conscious system has zero Φ and no experience, and a von Neumann simulation of a brain would not be conscious however perfect the simulation. David J. Chalmers notes the same architectural verdict for language models in "Could a Large Language Model Be Conscious?," *Boston Review*, August 9, 2023: on IIT, feed-forward transformer stacks lack integrated information. The convergence with Part Four is architectural, not doctrinal; the book's argument runs through content-fixing history rather than Φ.

**62.** Scott Aaronson, "Why I Am Not an Integrated Information Theorist (or, the Unconscious Expander)," *Shtetl-Optimized* (blog), May 2014, arguing that expander graphs and similar structures can be engineered to enormous Φ while doing nothing recognizable as experiencing; Tononi's reply, posted in the same exchange, accepts the consequence — a suitably organized grid would be conscious — as a prediction of the theory rather than a refutation. The exchange marks the joint precisely: IIT treats integration as sufficient in principle; the identity claim requires that integration organize *world-presenting content* before any question of phenomenal character arises.

**63.** Benjamin Libet, "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action," *Behavioral and Brain Sciences* 8, no. 4 (1985): 529–539, reports readiness potentials preceding reported awareness of the intention to move by roughly 350–400 ms. Three deflationary points. First, Libet himself defended a conscious "veto" and denied that his results abolish the control relevant to responsibility. Second, Aaron Schurger, Jacobo D. Sitt, and Stanislas Dehaene, "An Accumulator Model for Spontaneous Neural Activity Prior to Self-Initiated Movement," *Proceedings of the National Academy of Sciences* 109, no. 42 (2012): E2904–E2913, model the readiness potential as an artifact of averaging stochastic subthreshold fluctuations time-locked to movement onset. Third, Judy Trevena and Jeff Miller, "Brain Preparation Before a Voluntary Action: Evidence Against Unconscious Movement Initiation," *Consciousness and Cognition* 19, no. 1 (2010): 447–456, found comparable signals before a decision *not* to move. A paradigm of arbitrary, reasonless wrist-flexions models nothing about deliberation, which is the capacity actually at issue.

**64.** Immanuel Kant, *Critique of Pure Reason* (1781/1787), A444–451/B472–479 (the Third Antinomy), where transcendental freedom names a causality through which a series of appearances begins of its own accord, set against causality according to nature. Gilbert Ryle, *The Concept of Mind* (London: Hutchinson, 1949), ch. 1, for "the dogma of the Ghost in the Machine"; his ch. 3 treatment of the will anticipates much of the capacity-based analysis deployed here. The image of the ghost sitting "upstream" of the brain adapts Ryle's target to the free-will case.

**65.** The rigorous statement of the incompatibilist conditional is Peter van Inwagen's Consequence Argument (*An Essay on Free Will*, Oxford: Clarendon Press, 1983): if determinism is true, our acts follow from the laws of nature and the state of the world in the remote past, and neither is up to us. Crucially, the argument presses a modal claim about the actual past and the actual laws rather than a thesis about ordinary usage, so it is not met by observing what speakers mean by "could have done otherwise" — the reply in Section 18.5 concedes the modal point and shifts the ground. Van Inwagen does not grant the compatibilist an analysis of ability in exchange; he rejects the

conditional analyses on offer (*An Essay on Free Will*, ch. 4), which is why the reply here defends the capacity reading on its own merits rather than treating it as common ground. For the compatibilist rejoinder that "could have done otherwise" expresses an agent's ability rather than a physical-modal claim about identical universes, see David Lewis, "Are We Free to Break the Laws?," *Theoria* 47, no. 3 (1981): 113–121; the "new dispositionalism" of Kadri Vihvelin, Michael Fara, and Michael Smith develops the ability as a disposition whose manifestation may be determined away. Event-causal libertarianism (Robert Kane, *The Significance of Free Will*, New York: Oxford University Press, 1996) instead locates the required indeterminism in the neural processes of deliberation itself; its standing difficulty is the luck objection — an undetermined outcome seems attributable to chance rather than to the agent — pressed by, among others, Pereboom (n. 2).

**66.** Shaun Nichols and Joshua Knobe, "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions," *Noûs* 41, no. 4 (2007): 663–685, found that subjects presented with a deterministic scenario in abstract terms tended toward incompatibilist judgments, while subjects presented with a concrete, affect-laden case of the same determinism tended toward compatibilist ones. The upshot for the argument in the main text is that no bare appeal to "what ordinary people mean by 'could have done otherwise'" can settle the dispute, since the folk response varies with framing — which is why the chapter rests its reply on the excusing conditions internal to the practice of holding responsible rather than on a univocal folk intuition. Their own diagnosis runs the other way. Nichols and Knobe weigh two models of the split and lean, tentatively, toward the *performance-error* model: the underlying folk competence is incompatibilist, and the concrete, affect-laden cases distort its application ("when they are faced with a truly egregious violation of moral norms… they experience a strong affective reaction which makes them unable to apply the theory correctly"). The rival *affective-competence* model, on which the emotional response plays a constitutive rather than a distorting role, remains live, and the authors leave the choice unsettled. What the study establishes securely is the split itself. The chapter takes only that: the folk response varies with framing, so no bare appeal to what "we" mean settles the dispute — which is why the argument rests on the excusing conditions and on the ground for the capacity given in Section 18.3 rather than on any folk intuition, including one that would favor it.

**67.** Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, et al., "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness," arXiv:2308.08708v3 (2023), esp. 4–6 and secs. 1.2–2.5, doi:10.48550/arXiv.2308.08708. The report treats computational functionalism as a disputed working hypothesis and assigns credence according to functional similarity, support for the source theory, and confidence in the bridge. It explicitly warns that satisfying the indicators would not prove consciousness.

**68.** Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, et al., "Verbalizable Representations Form a Global Workspace in Language Models," arXiv:2607.15495v1, 16 July 2026, https://transformer-circuits.pub/2026/workspace/. The paper's abstract and secs. 1.1–1.5 support the functional claims stated here. The authors explicitly restrict their conclusion to access-consciousness-like organization and identify the missing clean analogues of recurrent loops, encapsulated modules, and sharp ignition.

**69.** Patrick Butlin, "Higher-order Representation in AI," *Philosophy and the Mind Sciences* 7 (2026), doi:10.33735/phimisci.2026.12032. Butlin's conclusion allows limited higher-order representation while retaining first-order, persona, and distality-based alternatives.

**70.** Simon Goldstein and Cameron Domenico Kirk-Giannini, "A Case for AI Consciousness: Language Agents and Global Workspace Theory," *Journal of Consciousness Studies* 33, nos. 7–8 (2026): 61–96; preprint arXiv:2410.11407. Their conclusion remains conditional on GWT and on a broadly computational-functionalist method.

**71.** David J. Chalmers, "Could a Large Language Model Be Conscious?," *Boston Review*, 9 August 2023; preprint arXiv:2303.07103. Chalmers treats recurrence, world and self models, global workspace, unified agency, senses, and embodiment as defeasible obstacles and engineering routes, while warning against mistaking performance for an operational definition of consciousness.

**72.** Patrick Butlin, Derek Shiller, and Jonathan A. Simon, "Introduction: Consciousness in Current AI," *Journal of Consciousness Studies* 33, nos. 7–8 (2026): 9–13. The introduction maps the special issue's disagreements over homeostasis, temporal continuity, GWT and language agents, LLM interpretation, and pluralist or natural-kind approaches.

**73.** Jeff Sebo and Robert Long, "Moral Consideration for AI Systems by 2030," *AI and Ethics* 5 (2025): 591–606, first published online 11 December 2023, doi:10.1007/s43681-023-00379-1. Their argument uses a non-negligible probability of consciousness as a threshold for moral consideration; the proportionate rule in the main text does not adopt their illustrative numerical threshold.

**73a.** A reader who reasons in expected values will want a number here. The proportional rule supplies the structure such a calculation needs without claiming its precision. The product of probability and stakes is only as good as the credence entering it, and that credence must come from bearer-specific evidence sorted by ledger. Fluency and uncorroborated self-report cannot carry it alone, and the bare conceivability of suffering carries none; reports that track interventions, and credible testimony about the system's organization, can contribute. A tiny, unevidenced credence multiplied by an enormous, scalable harm can dominate any calculation, which shows only that the inputs need disciplining. The rule therefore ties protection first to evidenced candidature and then to stakes.

**74.** For the relationalist contest, see Bill Brewer, "Perception and Content," *European Journal of Philosophy* 14, no. 2 (2006): 165–181, esp. 165–170, and *Perception and Its Objects* (Oxford: Oxford University Press, 2011); M. G. F. Martin, "The Transparency of Experience," *Mind & Language* 17 (2002): 376–425, https://doi.org/10.1111/1468-0017.00205, and "The Limits of Self-Awareness," *Philosophical Studies* 120 (2004): 37–89, esp. 38–41 and 67–72; and Charles Travis, "The Silence of the Senses," *Mind* 113, no. 449 (2004): 57–94, esp. 57–60 and 73–78. For John Campbell's demonstrative-reference argument and its relation to disjunctivism, see Matthew Soteriou, "The Disjunctive Theory of Perception," *Stanford Encyclopedia of Philosophy*, substantive revision July 28, 2026, secs. 3.5–3.6. These views agree that successful perception reaches the mind-independent object directly; they divide over whether content can capture that directness or the object must partly constitute the experience.

**75.** Ned Block, "Consciousness, Accessibility, and the Mesh between Psychology and Neuroscience," *Behavioral and Brain Sciences* 30 (2007): 481–548, esp. sec. 9, distinguishes phenomenal consciousness from access consciousness and uses partial-report paradigms associated with George Sperling and later work to argue that phenomenology overflows cognitive accessibility. The empirical interpretation remains disputed. Even if overflow defeats a strong access requirement, it does not by itself establish intrinsic mental paint; it may instead require a weaker account of availability or a different criterion for which representational states become conscious.

**76.** Global-workspace accounts treat availability as broadcast into a capacity-limited workspace; see Stanislas Dehaene, *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts* (New York: Viking, 2014). Recurrent-processing accounts locate the relevant transition earlier, in local recurrent perceptual activity; see Victor Lamme, "Towards a True Neural Stance on Consciousness," *Trends in Cognitive Sciences* 10, no. 11 (2006): 494–501. For higher-order theories, see Chapter 6, n. 7b. Patrick Butlin, Robert Long, et al., *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness* (2023), arXiv:2308.08708, secs. 2–3, derives indicators from several leading theories and treats their convergence as evidence rather than proof. That indicator method supports the epistemic policy in the text; it does not establish PANIC or collapse the competing mechanisms into one.

# The Figures — A Reader's Who's-Who

*Short orienting portraits of the philosophers the book engages — who they were, what they argued, and where they enter the story.*

### Thomas Aquinas

Thomas Aquinas (c. 1225–1274) was a Dominican friar, philosopher, and theologian whose writings bring Aristotelian philosophy into sustained discussion with Christian theology. His account of human nature combines bodily life and sensory cognition with an argument that intellectual activity has an immaterial principle. It therefore resists classification as either modern physicalism or Cartesian substance dualism.

### Major Works

- *Summa Theologiae* — a theological synthesis organized through questions, objections, and replies. Part I, questions 75–89, examines the human soul, its powers, and cognition.
- *Disputed Questions on the Soul* (*Quaestiones disputatae de anima*) — arguments about the soul's nature and relation to the body.
- *Commentary on Aristotle's De Anima* — interpretation of Aristotle's account of life, perception, and intellect.

### Soul and Body

Following Aristotle, Aquinas treats the soul as the **substantial form** of the living body: the principle in virtue of which the body is one living organism with its characteristic powers. "Form" here means more than shape, and the soul means more than conscious thought. Nutrition, sensation, and movement are among the activities of living beings that the doctrine explains.

The human being is a body–soul composite, not a soul using a body as an external instrument. Aquinas argues that sensation belongs to the living composite and requires bodily organs. He nonetheless holds that the rational soul can subsist apart from the body. These claims are intended to be compatible: a subsisting soul remains the form of a body, and its separated condition does not constitute complete human life.

### Perception and Intentional Reception

Aquinas distinguishes natural reception of a form from its cognitive or intentional reception. A colored thing can affect a sense organ without making the organ colored in the same way. Cognition receives a likeness or species through which the object becomes known; the knower need not become a material duplicate of the thing known.

This account belongs to the medieval background of theories of representation and intentionality. It should not be identified without qualification with a modern theory of neural vehicles and semantic content. Aquinas's explanation uses Aristotelian form, powers, and actuality, and different medieval thinkers developed different accounts of cognitive species.

### Abstraction and Intellectual Knowledge

Human understanding ordinarily begins from sensory experience. The internal senses provide **phantasms**, sensory images from which the active intellect abstracts intelligible species. Through

these the possible intellect comes to understand universal natures. Aquinas does not treat abstraction as extracting a physically detached object from an image; it makes an intelligible aspect available to intellectual cognition.

In *Summa Theologiae* I, question 85, article 2, he distinguishes what the intellect understands from that by which it understands. The intelligible species normally functions as a means of knowing the object, not as a substitute object that alone is known. Species can themselves become objects of reflection, but that is a further act. This distinction is central to discussions of whether his account is a direct or representational realism.

**The Argument for a Subsisting Intellect**

Aquinas argues that a power capable of knowing bodily natures universally cannot itself have the determinate bodily nature that would restrict its reception of them. He therefore holds that intellectual operation does not occur through a bodily organ in the way seeing does. From an operation belonging to the soul itself, he argues that the rational soul subsists; he also defends its incorruptibility.

This is an argument about the nature of intellectual cognition, not merely an appeal to revealed theology. Its premises about universality, bodily determination, and the dependence of operation on being are disputed. A contemporary critic may question whether universal thought requires an immaterial capacity at all. A Thomist replies that dependence on sensory and neural conditions does not by itself identify the intellectual act with an organ's operation; the philosophical issue concerns what that act can grasp.

**Interpretation and Continuing Disputes**

Two difficulties organize later discussion. First, how can the same soul function as the form of a body and subsist when separated from it? Aquinas's answer distinguishes complete human nature from the soul's capacity for independent existence, but interpreters disagree over the resulting account of personal identity. Second, do species explain knowledge of the external world or interpose representations between knower and object? The "by which" distinction supports a realist reading, while further questions remain about likeness, truth, and how cognitive forms relate to their objects.

Aquinas's body-dependent account of ordinary human thought also prevents the separability argument from being read as a claim that embodied life is dispensable. His psychology combines the rational soul's subsistence with the role of sensation and imagination in human understanding. Later Thomist interpretations emphasize different parts of that combination.

**Related Concepts and Figures**

- Aristotle — form, matter, and the powers of living things.
- *Franz Brentano* — a later recovery of Scholastic intentionality terminology.
- *Intentionality* — cognition directed toward its objects.
- *Embodiment* — bodily conditions of sensation and thought.
- *Dualism* — comparisons between hylomorphism and separable-mind theories.

**Further Reading**

- Aquinas, Thomas. *Summa Theologiae*, Part I, questions 75–89; especially question 75, articles 2 and 4, and question 85, articles 1 and 2.
- Aquinas, Thomas. *Disputed Questions on the Soul.*
- Aquinas, Thomas. *Commentary on Aristotle's De Anima.*
- Lagerlund, Henrik. "Mental Representation in Medieval Philosophy." *Stanford Encyclopedia of Philosophy.*

**Related Writing on This Site**

- *Mind, Matter, and Meaning* — Gordon Swobe's book-length argument about mind, consciousness, and meaning; an author position, not an endorsement by this reference page or its subjects.

**My View**

Aquinas belongs in the book as a limited precursor, not as the source or authority for its positive theory. His distinction between a thing's natural existence and the intentional presence of its form helps explain how cognition can concern the world without making the species a second object inspected by the knower. Brentano's later mark-of-the-mental thesis belongs to the same broad Aristotelian-Scholastic lineage, but it adds claims Aquinas did not make.

I keep that limited comparison and naturalize it. Content depends on world-involving causal history and producer-consumer organization, not on a special quasi-immaterial mode of existence. Aquinas supplies a useful historical resemblance about cognitive direction; he does not supply the book's identity claim, its theory of content, or its account of consciousness.

His subsisting intellect marks a real disagreement. The book keeps the embodied hylomorphic insight and rejects the claim that intellectual activity requires a separable power. That move pulls away from embodied mind, but it should not be presented as though Aquinas had already built the Cartesian inner theater. See *Vehicle and Content*, *Anti-Reification*, *Transparency*, and *Intentionality*.

**Aristotle**

Aristotle (384–322 BCE) studied under Plato for some twenty years, tutored the young Alexander the Great, and founded his own school at the Lyceum in Athens. His writings span logic, metaphysics, natural science, ethics, and psychology, and they shaped the intellectual traditions of the Islamic world and medieval Europe alike. In philosophy of mind, his *De Anima* ("On the Soul") remains the foundational text of naturalistic psychology and the starting point of any serious account of intentionality.

**Major Works**

- *De Anima* (c. 350 BCE) — the account of the soul as the form of the living body; the theory of perception as reception of form without matter
- *Parva Naturalia* — shorter works on perception, memory, dreaming, and related topics
- *Metaphysics* — the account of form and matter, act and potency, that underpins the psychology
- *Nicomachean Ethics* — practical reason and the embodied character of human flourishing

**Key Positions**

**Hylomorphism: the soul as the form of the body.** Aristotle's *De Anima* treats the soul (*psychē*) not as a thing housed in the body but as the *form* of a living body — its organization, its characteristic activity, what it is for that matter to be alive and to function. The soul is to the body roughly as cutting is to the axe, or sight to the eye. Mind, on this picture, is neither a separate substance nor mere matter but the organized, world-engaged activity of an embodied creature. Most psychic capacities are therefore powers of the living body and do not survive its destruction; whether Aristotle makes an exception for the active intellect remains a famously disputed question raised by *De Anima* III.5.

**Perception as reception of form without matter.** Aristotle's account of perception is that the sense faculty *receives the form of the sensible object without its matter*, as wax takes the imprint of a signet ring without the gold. The perceiver does not house an inner copy of the world; the perceiver takes on the world's form, is shaped by it, is *of* it. What that "reception" amounts to remains one of the longest-running disputes in Aristotle scholarship: Richard Sorabji reads it literally (the eye-jelly takes on color), Myles Burnyeat reads it as a purely intentional change with no ordinary physical alteration, and naturalistic readers treat it as the establishment of a real causal-informational relation between organism and world.

**Aristotle and the background to intentionality.** Aristotle did not formulate Brentano's modern mark-of-the-mental thesis, much less a general theory of diffuse intentional objects for moods and pains. His contribution lies further upstream: perception and thought take on the form of what they concern without taking on its matter. The Scholastics and Brentano developed that inheritance into explicit accounts of intentional directedness. Whether anxiety, pain, and mood possess content remains a further question, not an answer already supplied by Aristotle.

**The active intellect.** *De Anima* III.5 introduces the *nous poiētikos* (the active intellect) as "separable, impassive, unmixed." Readers have long taken the passage to posit a faculty that operates without any bodily organ and can survive the body's death. Aquinas and other medieval readers pulled on exactly this thread to make the rational soul detachable, and interpreters since antiquity have puzzled over how the passage squares with the hylomorphic picture, in which the soul is nothing apart from the living body.

**My View**

Aristotle offers the book a powerful pre-Cartesian comparison rather than a premise accepted on his authority. His hylomorphic picture shows that philosophy can describe mind as embodied, organized activity without placing a private spectator inside the body, and his account of perception provides part of the historical background to later theories of intentionality. Those affinities make him an important interlocutor in the book's genealogy, not the source from which its physicalism or identity claim follows.

I take Aristotle's *structure*, mind as embodied form and perception as world-contact, and decline his physics, his teleology of nature, and the substantive metaphysics of act and potency. The debt is architectural, not doctrinal. The *De Anima* III.5 passages on the active intellect are treated as a cautionary text, not a resource: the embodied-form picture is what to keep, and anything

posited as detachable from the body reopens the inner theater. See *Physicalism*, *Embodiment*, *Direct Realism*, and *Intentionality*.

### J. L. Austin

J. L. Austin (1911–1960) was a British philosopher associated with ordinary-language philosophy. His investigations of perception question the descriptions that generate philosophical arguments, while his philosophy of language examines what speakers do in saying something. Both projects treat differences among familiar uses of words as evidence that a theory must account for, not obstacles to be discarded in advance.

### Major Works

- *Sense and Sensibilia* (1962), reconstructed from manuscript notes by G. J. Warnock — perception and arguments for sense-data.
- *How to Do Things with Words* (1962), edited by J. O. Urmson — lectures on performatives and speech acts.
- "Other Minds" (1946) — knowledge claims and understanding other people.
- "A Plea for Excuses" (1956–1957) — distinctions among ways actions can go wrong and agents can be excused.

### Ordinary Language as a Starting Point

Austin examines how expressions function in different circumstances. Saying that someone acted accidentally, inadvertently, or under compulsion does not make the same claim three times. Similarly, saying that an object looks bent need not mean that it is bent or that another bent object is being perceived.

His method does not declare ordinary usage infallible. Ordinary language embodies distinctions developed through practical life and supplies an initial resource for identifying the phenomenon under discussion. Scientific discoveries and philosophical arguments may require new terminology. The objection is to theories that erase distinctions while continuing to rely on the intuitions those distinctions support.

### Perception and the Argument from Illusion

*Sense and Sensibilia* focuses on arguments associated with sense-datum theories, especially versions presented by A. J. Ayer. A familiar argument begins with a straight stick appearing bent in water. It infers that what is immediately perceived must be something bent other than the straight stick, and then generalizes to a sense-datum present in ordinary perception too.

Austin challenges both transitions. Seeing a straight stick under unusual viewing conditions need not mean failing to see that stick. The grammar of "looks bent" does not entail the existence of a second bent thing. Moreover, illusion, hallucination, distorted viewing conditions, mirrors, and misidentification are not interchangeable cases. A general theory cannot assume their sameness as a premise.

He also questions whether "direct" and "indirect" perception have a single philosophical contrast applicable to every example. Seeing a person in a mirror, seeing through spectacles, and seeing

someone through binoculars raise different questions. A theorist must specify the intended distinction rather than use "directly" to smuggle in an account of what perception is.

These criticisms do not themselves supply a complete positive theory. A defender of sense-data can formulate an argument about the properties apparently presented in experience without relying solely on ordinary grammar. Hallucination remains a problem for any account of perception. Austin's contribution constrains the arguments and descriptions; it does not establish that all competing theories have been refuted.

### "Real" as a Trouser-Word

Austin calls "real" a **trouser-word** because its negative use determines the relevant contrast. The negative, not the affirmative, "wears the trousers." A real duck might be contrasted with a toy, a decoy, a picture, or a stuffed specimen. Without knowing the contrast, one cannot assume that "real" attributes one uniform property.

The point is contextual, not a denial that some ducks exist independently of our judgments. It shows why broad questions about "real appearances" or "real objects" need clarification before they can support metaphysical conclusions. The terminology also corrects the tempting but inaccurate gloss that the affirmative word "real" does the decisive work in Austin's analysis.

### From Performatives to Speech Acts

Austin initially distinguishes **constative** utterances, which state something, from **performative** utterances, which accomplish something under suitable conditions. "I promise to return" can make a promise; the words do not merely report a pre-existing promise. Such acts have **felicity conditions**: the circumstances, authority, procedure, and sincerity required for them to succeed or avoid abuse.

As the lectures develop, Austin questions whether constatives and performatives form two cleanly separated classes. Statements also perform acts, and performatives can be assessed in ways related to truth and circumstances. He develops a more general analysis:

- **Locutionary act:** saying something with a particular sense and reference.
- **Illocutionary act:** what is done in saying it, such as asserting, warning, promising, or ordering.
- **Perlocutionary act:** what is brought about by saying it, such as persuading, alarming, or reassuring someone.

A warning can be issued without alarming its hearer; alarming someone can occur without a successful warning. Separating the act from its effects became a basis for subsequent speech-act theory, including John Searle's work.

### Criticism and Legacy

Critics question whether ordinary-language distinctions can decide questions about the underlying nature of perception, rather than merely our ways of talking about it. Defenders reply that a theory still needs accurate descriptions of its evidence. Speech-act theorists likewise debate the roles of convention, communicative intention, and uptake in successful illocutionary acts. Austin's work opens these questions rather than fixing one final taxonomy for all language use.

**Further Reading**

- Austin, J. L. *Sense and Sensibilia*. Reconstructed by G. J. Warnock. Clarendon Press, 1962.
- Austin, J. L. *How to Do Things with Words*. Edited by J. O. Urmson. Clarendon Press, 1962.
- Austin, J. L. "Other Minds." *Proceedings of the Aristotelian Society, Supplementary Volumes* 20 (1946): 148–187.
- Austin, J. L. "A Plea for Excuses." *Proceedings of the Aristotelian Society* 57 (1956–1957): 1–30.

**Related Writing on This Site**

- *Mind, Matter, and Meaning* — Gordon Swobe's book-length argument about mind, consciousness, and meaning; an author position, not an endorsement by this reference page or its subjects.

**My View**

Austin supplies one indispensable cut in the book's dismantling of the argument from illusion. He exposes the quick slide from *the stick looks bent* to *there is something bent of which I am aware*, and his distinction between illusion and delusion keeps a public object in view in the familiar stick case. That blocks the automatic inference to a private replacement.

It does not finish the argument by itself. A careful sense-datum theorist can grant that the straight stick remains present and still demand something that bears the presented bentness. At that point the Phenomenal Principle, not ordinary grammar alone, carries the burden. The book answers that slower argument through representational content and the vehicle/object distinction: a state can present bentness without anything bent existing as a second object.

Austin therefore remains an important interlocutor rather than the authority whose diagnosis settles perception. His analysis clears the verbal undergrowth; the positive account must still explain sensory presence, hallucination, content, and direct perceptual contact. See *Direct Realism* and *Indirect Realism and Sense Data*.

**Emily M. Bender and Alexander Koller**

Emily M. Bender and Alexander Koller are computational linguists whose joint paper "Climbing towards NLU" (2020) examines what success at language-processing tasks establishes about meaning and understanding. Their central distinction separates linguistic **form** from the relations between form, communicative intention, and the world. The paper challenges inferences from fluent output to understanding while arguing for more precise scientific descriptions of what systems learn.

**Relevant Works**

- Bender and Koller, "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data" (2020) — form, meaning, evaluation, and the octopus thought experiment.
- Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" (2021) — risks of large language models, including environmental cost, data practices, bias, and misleading fluency.

The second paper is not a joint work by Bender and Koller. Its coauthors and broader concerns should be distinguished from the 2020 argument rather than used to assign one unchanging position to both researchers.

### Form and Meaning

For Bender and Koller, form includes observable linguistic signals and their distribution: written strings, spoken sounds, and patterns of combination. Meaning involves a relation between these forms and something beyond them, including the speaker's communicative intent. Predicting a continuation of a sentence and understanding what a speaker is trying to convey are consequently different achievements.

Their claim is conditional: a system trained **only on form** has no way to learn meaning in the sense they specify. It is not the claim that language models learn nothing, that statistical methods have no scientific value, or that no artificial system could understand. The challenge concerns what an objective based on linguistic form can establish without access to the relevant nonlinguistic relations.

The authors explicitly distinguish form-only learning from some semantically supervised text tasks. Reading-comprehension datasets can specify semantic relations between pieces of text; they allow that a sufficiently sophisticated model trained on such data might learn aspects of meaning. Pretrained language-model representations might help that downstream learning without themselves constituting meaning. Thus **text-only** need not mean **form-only**. The concession does not establish full understanding, and the authors warn that task success can still exploit statistical artifacts.

### The Octopus Thought Experiment

Two people on separate islands communicate through a telegraph cable. An octopus, able to intercept the signals but unable to observe the islanders' world, learns regularities in their exchanges. It eventually imitates one correspondent well enough to deceive the other.

The difficulty becomes concrete when an islander needs help dealing with a bear using materials at hand. Appropriate advice requires identifying objects, interpreting the predicament, and relating an action to its likely consequences. Bender and Koller argue that the octopus's signal-pattern learning does not supply those relations. A listener may nevertheless interpret its output as meaningful, attributing intentions that the octopus has not acquired.

The argument is not that a fluent system must give an unhelpful answer to every practical question. A memorized or well-predicted message might happen to help. The issue is whether a system has learned the meaning relation, rather than whether a particular output can be useful to someone who already understands it.

### Why the Argument Is Contested

One response holds that language is itself produced by world-involving agents, so texts can transmit information about objects, situations, and social practices. Language-only learning may extract substantial structure from this evidence. The question then becomes whether learning correlations projected into text yields the relations constitutive of meaning, or only a model of their linguistic consequences.

Another response disputes the authors' definition of meaning or adopts a functional criterion of understanding. On such a view, systematic success at inference, explanation, and appropriate

response can count toward understanding even without human-like experience or embodiment. Bender and Koller's challenge remains that criteria and evidence must be stated independently of the fluent behavior they are supposed to assess.

Multimodal training, interaction, tool use, and feedback can also change the conditions of the thought experiment. The relevant question is which relations the system can learn, not merely whether additional data are present. Bender and Koller consider a system given text and an unrelated collection of images: those two sources alone do not supply the missing connection between an utterance and its pictured referent. Whether paired data or interaction supply grounding, communicative intention, or robust understanding needs separate argument and evaluation. Neither calling them "grounding" nor continuing to call the system an "octopus" settles their effects.

### Stochastic Parrots and Broader Risks

The 2021 paper examines how fluent text generation can invite readers to attribute coherence, factual authority, and communicative intent. It also addresses costs and harms that do not depend on a final theory of machine understanding: energy use, poorly documented training data, reproduction of social biases, and attention diverted from affected communities.

These are distinct claims. A system could have substantial language competence while remaining environmentally costly or unsafe. Conversely, reducing a deployment's harms would not prove that the system understands. Keeping the semantic and social-risk arguments separate makes both easier to assess.

### Relation to the Chinese Room

The octopus and John Searle's Chinese Room both question whether linguistic performance establishes understanding. Their mechanisms and targets differ. The octopus learns distributions from data and lacks access to grounding relations; the Chinese Room asks whether formal program execution suffices for understanding. Accepting one argument does not automatically require Searle's account of consciousness or biology, and rejecting a simple analogy between the cases does not answer either argument on its own.

### Further Reading

- Bender, Emily M., and Alexander Koller. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020): 5185–5198. doi:10.18653/v1/2020.acl-main.463.
- Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*: 610–623. doi:10.1145/3442188.3445922.

### Related Writing on This Site

- *Mind, Matter, and Meaning* — Gordon Swobe's book-length argument about mind, consciousness, and meaning; an author position, not an endorsement by this reference page or its subjects.

### My View

Bender and Koller are allies, and I build directly on them. Their octopus is the cleanest recent demonstration that linguistic form alone cannot fix meaning, distal reference, or communicative intent. The book borrows it without apology and separates the achievements their paper leaves open. Public practice can transmit derived content and inherited reference; training can make that content causally operative inside learned mechanisms; world-involving selection and use can establish original content where they fix the semantic standard independently. None follows from distributional form by itself, and none alone establishes understanding by a whole system.

The categorical point concerns *form alone*, not every system trained on text or every capacity of such a system. Scale and more text cannot make internal distribution fix a distal address. Corpus-mediated history, multimodal contact, tools, persistent control, and embodied interaction may add relations the octopus lacks; each must be assessed rather than smuggled in through fluency. Bender and Koller gave the field its indispensable thought experiment. The book uses it as a limit on what form establishes, not as a permanent verdict on artificial understanding.

### Ned Block

Ned Block is an American philosopher whose work concerns consciousness, perception, and the foundations of cognitive science. He distinguishes phenomenal experience from cognitive access and challenges accounts that identify experience entirely with functional organization or representational content. These criticisms operate within debates informed by psychology and neuroscience; rejecting a functional or representational theory does not by itself amount to rejecting physicalism.

### Major Works

- "Troubles with Functionalism" (1978) — absent and inverted qualia as challenges to functional accounts of mind.
- "Inverted Earth" (1990) — phenomenal character and environmentally individuated content.
- "On a Confusion about a Function of Consciousness" (1995) — phenomenal and access consciousness.
- "Mental Paint" (2003) — qualitative character, introspection, and representationalism.
- "Consciousness, Accessibility, and the Mesh between Psychology and Neuroscience" (2007) — evidence concerning experience and cognitive access.

### Phenomenal and Access Consciousness

**Phenomenal consciousness**, or P-consciousness, concerns what it is like to undergo an experience: the felt character of a pain or the visual appearance of a scene. **Access consciousness**, or A-consciousness, concerns a state's content being available for reasoning and the rational control of action and speech. Availability need not involve an actual report at the moment the state occurs.

Block argues that treating these as interchangeable can distort explanations of consciousness. A mechanism explaining reportability may explain access without thereby explaining phenomenal character. The distinction itself leaves open whether the two depend on the same processes in actual human beings.

His hypothetical **super-blindsighter** illustrates access without experience: visual information becomes available for use in reasoning and behavior although the subject has no visual phenomenology. This is an imagined extension of blindsight, not a description established for ordinary blindsight patients. In the other direction, Block argues that the richness of perceptual experience can exceed what limited-capacity cognition can access.

### Absent and Inverted Qualia

Functionalism individuates mental states by their roles in a system of inputs, internal transitions, and outputs. Block asks whether the same roles could be realized without experience or with systematically different experience. If so, those roles would not suffice to identify phenomenal states.

The large-scale system in "Troubles with Functionalism" arranges people and communication devices to reproduce a functional organization resembling that of a human mind. The question is whether the resulting system has experience. The argument is not that every unfamiliar substrate is incapable of mentality; it tests whether functional equivalence alone guarantees it.

Functionalists reply that intuitions about such systems are unreliable or concern the parts rather than the whole. Some accept that the appropriately organized system would be conscious. Others specify a richer functional role or the relevant level of organization. Block's objection then turns on whether the revised specification explains consciousness independently or builds the disputed property into its definition.

### Mental Paint and Representationalism

Representationalists hold, in different strengths, that phenomenal character depends on or is identical with how experience represents the world. Block distinguishes **mental paint** from **mental oil**. Paint consists of phenomenal properties that do represent, but whose nature, he argues, is not exhausted by their representational content. Oil consists of phenomenal properties that do not represent anything. Neither metaphor requires a literal inner picture or a second object watched by a mental spectator.

His cases include:

- **Bodily sensations:** Block proposes that experiences such as orgasm include mental oil. He allows that a sensation can have content while arguing that some of its phenomenal features represent nothing. Representationalists contest that diagnosis and offer richer accounts of bodily content.
- **Phosphenes and afterimages:** visual qualities can occur without the corresponding external objects. Block's phosphene objection goes beyond that absence: even granting a representation of colored moving expanses, he claims awareness of experiential features beyond that content. The disputed point concerns introspective access, not simply whether the apparent object exists.
- **Inverted Earth:** environmental and linguistic changes, paired with compensating sensory changes, test whether externally individuated content and phenomenal character can vary independently. Block presents this as an argument for paint, not oil: representing something need not exhaust the representing experience's nature.

Representationalists reply that phosphenes can represent colors and locations inaccurately rather than introduce nonrepresentational qualities. A related debate asks whether seeing a sharp edge blurrily differs in content from seeing a fuzzy edge clearly; one response distinguishes the determinacy of represented spatial information. Mode-sensitive accounts also distinguish ways of presenting content. The resulting dispute concerns whether these distinctions are independently justified and capture the examples, or whether "content" is being expanded solely to absorb counterexamples.

Transparency is also contested. The claim that introspection directs attention toward represented features is weaker than the claim that there are no introspectible properties beyond representational content. Block's cases target the stronger inference, not merely the fact that ordinary visual attention concerns the scene.

**Overflow and Empirical Evidence**

Block's later arguments connect the phenomenal/access distinction with experiments on partial report, visual memory, and neural processing. Subjects can seem to experience a rich display while reporting only a few items. He argues that phenomenal representation may overflow the capacity of cognitive access.

Critics distinguish conscious detail from unconscious information preserved briefly and accessed after a cue. They also question whether an impression of richness shows that every detail was consciously represented. Block replies that converging psychological and neural evidence can support a separation even when immediate report is an incomplete measure. The methodological problem is genuine: testing inaccessible experience cannot simply rely on asking subjects to report it. No single partial-report result settles the debate.

**Further Reading**


- Block, Ned. "Troubles with Functionalism." *Minnesota Studies in the Philosophy of Science* 9 (1978): 261–325.
- Block, Ned. "Inverted Earth." *Philosophical Perspectives* 4 (1990): 53–79.
- Block, Ned. "On a Confusion about a Function of Consciousness." *Behavioral and Brain Sciences* 18 (1995): 227–247.
- Block, Ned. "Mental Paint." In *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, edited by Martin Hahn and Bjørn Ramberg, 165–200. MIT Press, 2003.
- Block, Ned. "Consciousness, Accessibility, and the Mesh between Psychology and Neuroscience." *Behavioral and Brain Sciences* 30 (2007): 481–499.


**Related Writing on This Site**

- *Mind, Matter, and Meaning* — Gordon Swobe's book-length argument about mind, consciousness, and meaning; an author position, not an endorsement by this reference page or its subjects.

**My View**

Ned Block is the book's most formidable opponent on its own turf. Where Searle attacks computationalism from outside and the dualist attacks physicalism from outside, Block attacks the book's *positive* theory, strong representationalism, from inside the broadly physicalist, scientifically

serious camp the book itself occupies. He is a naturalist, no friend of dualism, and he still thinks the central identity claim is false. That makes him the opponent worth meeting at his best.

The book uses Block's P/A distinction while resisting the moral he draws from it — that P-consciousness floats free of representational facts in a way that defeats representationalism. On my view, the "what it is like" is itself representational; the distinction marks two roles, not two kinds of stuff.

On mental paint, I answer blur through determinacy and manner: blurry vision presents edges less determinately rather than revealing a smear on an inner canvas. Afterimages and phosphenes press harder. A nonconceptual visual state may present a colored, filmy shape at an apparent location while the subject's higher-level judgment correctly identifies the episode as an afterimage or phosphene. That reply remains contested rather than settled, and transparency alone cannot establish the identity claim. Block is the reason the book must specify content, mode, attention, and state-consciousness independently rather than enlarging "content" whenever a difficult case appears. The pressure is real; the further mental paint still explains nothing the complete intentional profile does not explain better. See *Transparency*, *Qualia*, and *Michael Tye*.

**Franz Brentano**

Franz Brentano (1838–1917) was a philosopher and psychologist whose account of mental acts helped shape later phenomenology and analytic philosophy of mind. In *Psychology from an Empirical Standpoint* (1874), he characterizes mental phenomena through their directedness toward objects. His terminology draws on Aristotle and Scholastic philosophy, while his project seeks a systematic, empirically grounded study of mental life.

**Major Works**

- *Psychology from an Empirical Standpoint* (1874) — mental phenomena, intentionality, inner perception, and their classification.
- *On the Origin of Our Knowledge of Right and Wrong* (1889) — judgment and evaluative attitudes in moral knowledge.
- *Descriptive Psychology* — lectures published posthumously in German in 1982 and in English translation in 1995.

**Intentionality as the Mark of the Mental**

Brentano proposes that every mental phenomenon involves reference to a content or direction toward an object. Hearing concerns a sound, judging concerns what is affirmed or denied, and loving concerns something loved. "Intentional" does not here mean deliberate: a perception can be intentional without having been chosen.

He contrasts this with **physical phenomena** as he uses that expression, such as a color or sound presented in experience. It is misleading to substitute a modern contrast between conscious brains and every physical state in the universe without qualification. His classification and his account of what external perception warrants do not coincide with a contemporary scientific ontology.

Later versions of the mark-of-the-mental thesis ask both whether every mental state has intentionality and whether only mental states do. Pains, moods, and undirected anxiety challenge

the first claim. Representational accounts of nonconscious information-processing systems raise questions about the second. These are developments of Brentano's problem, not simple experimental tests of his original vocabulary.

### Intentional Inexistence

The phrase **intentional inexistence** concerns an object's presence in or involvement with a mental act. It does not mean simply that the object does not exist. Thinking of an imaginary creature and perceiving an actual tree both exhibit directedness, even though only one has an ordinary existing object.

Brentano's early language of immanence invited interpretations involving an object internal to the act. Scholars disagree about the ontological commitments of this formulation and its relation to his later views. It should not be treated as an uncomplicated claim that every perception is awareness of a miniature object inside the head. Equally, his terminology cannot simply be replaced by contemporary externalist content without acknowledging the interpretive problem.

The difficulty becomes acute for nonexistent objects. If thinking is a relation, what stands at the other end when someone thinks of a golden mountain? Different members of Brentano's school developed different answers. Meinong's object theory, Husserl's analyses of intentional consciousness, and Twardowski's distinctions among act, content, and object are divergent developments, not doctrines all shared by Brentano.

### Presentations, Judgments, and Love or Hate

Brentano divides mental phenomena into three basic classes:

- **Presentations:** having something before the mind, whether through perception or imagination.
- **Judgments:** accepting or rejecting what is presented.
- **Phenomena of love and hate:** positive and negative evaluative attitudes, including forms of desire and emotion.

Judgment is not, on his account, just combining concepts into a sentence. A presented object can be accepted or rejected. Likewise, an evaluative attitude differs from merely presenting its object. The classification focuses on different kinds of act, rather than treating mental life as a container of passive ideas.

### Inner Perception and Consciousness

Brentano distinguishes **inner perception** from deliberate inner observation. A person hearing a sound is primarily directed toward that sound, while the act also involves awareness of itself as a secondary object. This does not require an additional observer or an indefinitely repeated series of monitoring acts.

He takes inner perception to provide immediate evidence unavailable in the same way from external perception. His account has influenced discussions of self-awareness and the unity of consciousness. Critics question its claim to evidential security and its apparent exclusion of

unconscious mental phenomena. Later theories of intentionality often retain directedness without accepting Brentano's account of consciousness or certainty.

### Development and Later Reism

Brentano's ontology changed over his career. In his later **reism**, he rejects abstract entities and other non-thing objects in favor of individual things, and attempts to reformulate apparently committing expressions. This development complicates readings that project the early language of immanent objects unchanged across his entire philosophy.

His **descriptive psychology** analyzes the components and structures of mental phenomena, whereas genetic psychology concerns their causal conditions. The distinction separates description of a mental act from an explanation of how it arises. It does not imply that psychological descriptions and causal investigations are mutually exclusive.

### Objections and Continuing Questions

Contemporary intentionalists can respond to apparent undirected states by identifying bodily conditions, diffuse situations, or the world more generally as what those states concern. Critics ask whether that stretches directedness beyond explanatory usefulness. Similarly, attributing content to a detector need not attribute beliefs or consciousness to the whole device. The scope of intentionality, the identity of its bearer, and its relation to conscious experience remain distinct questions.

### Related Concepts and Figures

- *Intentionality* — the structure and scope of directedness.
- *Transparency* — a later question about what introspection reveals.
- Aristotle — historical sources of Brentano's psychology.
- *Thomas Aquinas* — Scholastic cognition and intentional terminology.
- *Tim Crane* — contemporary defenses of intentionality as a mental mark.

### Further Reading

- Brentano, Franz. *Psychology from an Empirical Standpoint*. Originally published 1874. Translated by Antos C. Rancurello, D. B. Terrell, and Linda L. McAlister; English edition edited by McAlister. Routledge, 1995.
- Brentano, Franz. *Descriptive Psychology*. Translated and edited by Benito Müller. Routledge, 1995.
- Huemer, Wolfgang. "Franz Brentano." *Stanford Encyclopedia of Philosophy*. Substantive revision 2019.

### Related Writing on This Site

- *Mind, Matter, and Meaning* — Gordon Swobe's book-length argument about mind, consciousness, and meaning; an author position, not an endorsement by this reference page or its subjects.

### My View

Brentano supplies the book's starting question, not its final threshold. His core insight survives: mental states have content and are individuated partly by what they present. But finding content

in a detector, perceptual subsystem, or trained mechanism does not by itself establish a whole mind. Which persisting organization believes, understands, acts, or is conscious by means of that state remains a separate question about integration, control, memory, correction, planning, and phenomenal organization.

The difficulty is the immanent-object reading. Brentano's language of intentional inexistence, the object existing *within* the act, naturally generates the picture of a mind that carries around a private theater of inner objects, a mental replica of the world enclosed in consciousness. This is exactly the picture the book dismantles. Harman's transparency argument, the Tye/Dretske identity claim, Austin's demolition of sense-datum theory — all of these can be read as the sustained effort to honor Brentano's intentionality insight while rejecting the immanent-object reading.

The correction is now standard: the core insight (directedness as the mark of the mental) survives; the ontology of immanent objects does not. The object of perception is the world, not a mental surrogate. This is precisely the move from Brentano's act psychology to a properly externalist intentionalism. See *Intentionality*.

**Tyler Burge**

Tyler Burge is an American philosopher whose work examines how thought and perception depend on relations beyond the individual. His **anti-individualism** holds that some mental contents are constitutively shaped by social practices or environmental relations. It does not merely observe that the world causes mental states: it claims that what a state represents can depend on more than an individual's intrinsic physical and psychological condition.

**Major Works**

- "Individualism and the Mental" (1979) — social conditions on thought, developed through the arthritis case.
- "Individualism and Self-Knowledge" (1988) — compatibility between anti-individualism and first-person knowledge.
- "Perceptual Entitlement" (2003) — warrant for perceptual belief without a prior argument establishing reliability.
- *Origins of Objectivity* (2010) — perceptual representation, constancies, and the origins of objective reference.

**The Arthritis Thought Experiment**

Burge imagines a person who uses "arthritis" in largely ordinary ways but incorrectly believes that arthritis can occur in the thigh. The person is willing to accept correction by medical experts. In the actual linguistic community, arthritis concerns joints, so the expressed belief about arthritis in the thigh is false.

Now consider a counterfactual community that uses the same word for a wider range of rheumatoid ailments, including ones outside the joints. Hold fixed the subject's relevant internal history and dispositions while varying the surrounding linguistic practice. Burge argues that the subject's

thought then has a different content: it concerns the wider condition, rather than arthritis as understood in the original community.

The example depends on incomplete understanding, not complete ignorance. People can possess a concept and have beliefs involving it without mastering every application. Social correction and deference help explain why the actual subject's error is a false belief about arthritis, rather than a private redefinition that makes the statement automatically correct.

### What Anti-Individualism Claims

The same intrinsic individual condition can occur with different contents in different social or environmental settings. This is a claim about **individuation**—what makes a thought the thought it is. It should not be paraphrased as an instantaneous change in brain activity whenever a community changes a word.

Nor does it mean that every person must defer to experts about every concept. The force of the example depends on the subject's participation in a practice and the particular counterfactual conditions. Anti-individualism has social and perceptual forms; social membership is not a necessary condition for every kind of representation.

Critics propose a **narrow content** shared by internally identical subjects or argue that the arthritis case concerns linguistic interpretation rather than thought itself. Burge replies that ordinary psychological explanations attribute contents involving shared concepts and that restricting content to what is fixed internally omits part of the phenomenon. Whether a narrower description can serve some scientific purposes is a further issue, not automatically a refutation of wide content.

### Perceptual Representation and Objectivity

In *Origins of Objectivity*, Burge investigates how perception represents objective features of the environment. Perceptual **constancies** are central: a system can represent a stable shape, size, or color despite changes in proximal stimulation due to distance, viewpoint, or illumination. The resulting state is not merely a record of what is currently happening at a receptor.

Such representation has veridicality conditions and can be accurate or inaccurate. Its individuation depends partly on the perceptual system's relations to environmental kinds and on its history. Mere causal sensitivity is not automatically perception; a detector reacting to a signal does not necessarily exercise the representational capacities Burge analyzes.

Objective perception does not require the perceiver to possess philosophical concepts of mind-independent objects or to reason about its own reliability. Nonhuman animals can possess perceptual representation without those reflective abilities. Burge opposes theories that impose sophisticated conceptual prerequisites on basic perception.

### Self-Knowledge and Perceptual Entitlement

Anti-individualism appears to create a puzzle about self-knowledge. If content depends on surroundings, must a person investigate those surroundings before knowing what they think? Burge argues that immediate knowledge of one's present thought can remain authoritative even

when one lacks complete knowledge of the conditions that individuate it. Knowing that one thinks something and independently identifying its entire constitutive background are different achievements.

His account of **perceptual entitlement** similarly distinguishes warrant from a subject's ability to provide a reflective justification. A perceiver can have warranted beliefs based on a reliably functioning perceptual system without first proving the system's reliability. The entitlement is defeasible: contrary evidence can undermine it. It is not a guarantee that every perception or belief is accurate.

### Development and Criticism

Burge's work extends from social conditions on conceptual thought to biological and psychological conditions on perceptual objectivity. The later project does not retract the earlier externalism; it broadens the kinds of environmental dependence under investigation.

Debates concern the role of history, the individuation of content in unusual environments, and the compatibility of wide content with causal explanation. Philosophers also ask whether self-knowledge combined with externalism permits unexpectedly easy knowledge of the external world. The theories require distinctions among possessing a concept, applying it correctly, knowing one's thought, and knowing the environment. Anti-individualism by itself is not a complete theory of phenomenal consciousness or artificial understanding.

### Further Reading

- Burge, Tyler. "Individualism and the Mental." *Midwest Studies in Philosophy* 4 (1979): 73–121.
- Burge, Tyler. "Individualism and Self-Knowledge." *The Journal of Philosophy* 85 (1988): 649–663.
- Burge, Tyler. "Perceptual Entitlement." *Philosophy and Phenomenological Research* 67 (2003): 503–548.
- Burge, Tyler. *Origins of Objectivity*. Oxford University Press, 2010.

### Related Writing on This Site

- *Mind, Matter, and Meaning* — Gordon Swobe's book-length argument about mind, consciousness, and meaning; an author position, not an endorsement by this reference page or its subjects.

### My View

Burge is a crucial ally, and *Origins of Objectivity* stands behind more of this book than the citations show. His account of perception as fundamentally world-directed, objectively individuated, and anti-individualistically constituted converges with the book's core claim that experience presents the world, not inner surrogates.

The social externalism is exactly right, and the book inherits it: content is not in the head; mental states are individuated by their relations to the world and the community. This is part of what makes the book's qualified reference theory of meaning possible — reference is an achievement that depends on social and environmental relations, not on matching inner representations to outer facts.

The departures: Burge tends to treat phenomenal consciousness at arm's length — he is careful about the veridicality conditions of perceptual states, but less engaged with what it is *like* to be in those states. The book carries anti-individualism into the identity claim without identifying phenomenal character with bare world-directed content. Within a state already counted as conscious and borne by an identified subject, phenomenal character consists in the complete world-presenting intentional profile: externally fixed content under a perceptual or bodily-affective mode, with its determinacy, field structure, and attentional organization.

One further step concerns what follows from objective content. Burge's evolutionary and perceptual history helps explain how states acquire veridicality conditions. The book then asks separately whether such content constitutes phenomenal character and which integrated organization perceives or understands by means of it. Practical stake may add autonomy, vulnerability, and welfare, but it neither fixes the content nor creates objectivity. See *Semantic Externalism* and *Hilary Putnam*.

**David Chalmers**

David J. Chalmers is an Australian philosopher whose work concerns consciousness, modal reasoning, meaning, and virtual reality. His distinction between the easy and hard problems of consciousness organizes a dispute about what physical and functional explanations can establish. He defends arguments against standard materialism while developing proposals for studying consciousness as part of the natural world.

**Major Works**

- "Facing Up to the Problem of Consciousness" (1995) — the easy/hard problem distinction.
- *The Conscious Mind: In Search of a Fundamental Theory* (1996) — arguments against reductive explanation and a framework for naturalistic dualism.
- "Does Conceivability Entail Possibility?" (2002) — distinctions among kinds of conceivability and possibility.
- "Consciousness and Its Place in Nature" (2002/2003) — a taxonomy of approaches to the metaphysics of consciousness.
- "The Meta-Problem of Consciousness" (2018) — explaining why people make judgments about consciousness and its apparent mystery.
- *Reality+: Virtual Worlds and the Problems of Philosophy* (2022) — virtual objects, simulation, knowledge, and value.

**Easy and Hard Problems**

The **easy problems** concern capacities such as discrimination, information integration, attention, and report. They are not necessarily easy to solve; Chalmers calls them easy because explaining the appropriate mechanisms and functions would explain the capacity under investigation.

The **hard problem** asks why those processes are accompanied by phenomenal experience: why there is something it is like to undergo them. Chalmers argues that a complete functional account can leave this question unanswered. This is a proposed distinction between explanatory tasks, not a claim that neuroscience has no bearing on experience.

Critics maintain that the distinction assumes what a successful physical account must explain, or that progress in explaining cognitive organization will remove the apparent residual question. Chalmers replies that an explanation of discrimination and report remains different from an explanation of why discrimination and report involve feeling. The disagreement concerns explanatory sufficiency, not the value of empirical research.

### Zombies and the Modal Argument

A philosophical zombie is a physical duplicate without phenomenal consciousness. Chalmers argues that an ideally coherent zombie scenario is conceivable and uses principles of modal epistemology to connect this with possibility. If all the physical facts could hold without the phenomenal facts, the former would not metaphysically necessitate the latter as standard physicalism requires.

This is not simply an inference from what someone can picture to what nature permits. His argument distinguishes ideal from initial conceivability, positive scenario construction from absence of a detected contradiction, and primary from secondary evaluation. Nor does a merely behavioral or functional duplicate establish the same conclusion: such a duplicate need not preserve all physical facts.

Physicalists can dispute the conceivability premise, its modal implications, or the account of phenomenal concepts. Type-B physicalists accept an epistemic gap while defending a necessary a posteriori identity. Phenomenal-concept strategies explain the gap through distinctive first-person ways of thinking about the same physical property. Chalmers argues that this response must still justify the necessary connection rather than merely explain cognitive independence.

### Two-Dimensional Semantics

Chalmers's epistemic two-dimensional framework distinguishes two ways of evaluating a scenario:

- A **primary intension** evaluates a scenario considered as actual: what an expression would apply to if that scenario described how the world actually is.
- A **secondary intension** evaluates a counterfactual world while retaining the reference fixed by the actual world.

A primary intension is therefore not simply an expression's reference in the current context of utterance. Both intensions concern systematic evaluation across scenarios, but their roles differ.

For example, a scenario in which the clear drinkable liquid is XYZ can, considered as actual, verify that water is XYZ. Once actual water is identified as $H_2O$, a counterfactual XYZ liquid is not water. This explains how an empirically discovered identity can be necessary without being a priori.

Chalmers argues that phenomenal concepts do not admit the same separation between apparent and underlying reference: a genuinely pain-like experience is already a pain. The argument therefore needs further steps connecting primary zombie conceivability with the relevant metaphysical conclusion. Critics dispute those steps, his treatment of phenomenal concepts, or the framework itself. The sophisticated argument also leaves a role for Russellian views of the intrinsic nature

underlying physical structure; it should not be compressed into "primary conceivability always entails secondary possibility."

### Naturalistic Dualism and Russellian Alternatives

**Naturalistic dualism** treats phenomenal properties as irreducible while connecting them to physical processes through psychophysical laws. It need not posit an independently acting immaterial substance. Chalmers also investigates panpsychist and panprotopsychist alternatives: respectively, views assigning fundamental experience or experience-grounding properties to basic reality.

These are not interchangeable theories or a simple narrative of replacing dualism with panpsychism. Russellian proposals distinguish the structure described by physics from its intrinsic categorical basis. Their difficulties include the **combination problem**: explaining how fundamental experiences or properties yield the unified experience of a person. Chalmers treats that problem as a substantive obstacle, not something resolved by making experience widespread.

### The Meta-Problem and Virtual Reality

The meta-problem asks why we judge that consciousness presents a hard problem. Because these judgments and reports are behaviors and cognitive processes, Chalmers treats their explanation as functionally tractable. Illusionists argue that such an explanation can remove the need for the hard problem. Chalmers does not accept that conclusion; he argues that a theory must explain the judgments as well as what they concern.

In *Reality+*, he argues that virtual objects can be real digital objects and that life in a virtual world can have genuine value. This does not mean every virtual claim is true or every simulation contains a conscious subject. Questions about the reality of virtual objects, knowledge of one's environment, and the consciousness of simulated agents require separate treatment.

### Further Reading


- Chalmers, David J. "Facing Up to the Problem of Consciousness." *Journal of Consciousness Studies* 2 (1995): 200–219.
- Chalmers, David J. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, 1996.
- Chalmers, David J. "Does Conceivability Entail Possibility?" In *Conceivability and Possibility*, edited by Tamar Szabó Gendler and John Hawthorne, 145–200. Oxford University Press, 2002.
- Chalmers, David J. "Consciousness and Its Place in Nature." In *The Blackwell Guide to Philosophy of Mind*, edited by Stephen P. Stich and Ted A. Warfield. Blackwell, 2003. Also published in *Philosophy of Mind: Classical and Contemporary Readings*, edited by Chalmers, Oxford University Press, 2002.
- Chalmers, David J. "The Meta-Problem of Consciousness." *Journal of Consciousness Studies* 25, nos. 9–10 (2018): 6–61.
- Chalmers, David J. *Reality+: Virtual Worlds and the Problems of Philosophy*. W. W. Norton, 2022.

**Related Writing on This Site**

- *Mind, Matter, and Meaning* — Gordon Swobe's book-length argument about mind, consciousness, and meaning; an author position, not an endorsement by this reference page or its subjects.

**My View**

Chalmers is the philosopher the book takes most seriously as an opponent — which means crediting him before resisting him. The hard problem framing is genuinely illuminating: something does need explaining that functional and access accounts leave out. The explanatory gap is real. I am with Chalmers that far.

The departure is on what follows. Chalmers moves from the undeniable epistemic gap (we cannot derive phenomenal facts from physical descriptions) to an ontological conclusion: phenomenal properties must therefore be metaphysically distinct. This slide from epistemic to ontological is exactly where the book's argument intervenes. A gap in our conceptual apparatus is not automatically a gap in the world. We cannot derive phenomenal descriptions from physical ones because we hold phenomenal concepts in a special first-person way that doesn't reveal their physical-representational basis — not because phenomenal properties float free of physical reality. The identity claim (phenomenal character *consists in* representational content of the right embodied, world-directed kind) is the positive story about why the gap exists without licensing Chalmers' dualist conclusion.

On zombies: I grant Chalmers the conceivability without reservation, and think he has been right to insist that no amount of neuroscience will dissolve it. Our phenomenal concepts pick out physical-representational facts by a route that runs through undergoing rather than describing, and two concepts related that way stay cognitively independent whatever they refer to. That independence is what the conceiving registers. Chalmers reads it as a window onto the space of possible worlds; I read it as the permanent shadow of one fact approached along two paths — which is exactly what an identity of this kind should be expected to cast.

Chalmers' panpsychist turn makes things worse, not better. Positing fundamental phenomenal properties at the micro-level simply relocates the explanatory gap without closing it. See *The Hard Problem of Consciousness* and *Physicalism*.

**Alonzo Church**

Alonzo Church (1903–1995) was an American logician and mathematician whose work helped establish computability theory. His lambda calculus, undecidability results, and theory of types provide formal tools for analyzing functions, proof, and effective procedures. His contributions distinguish questions that have algorithms from questions for which no general algorithm exists; they do not by themselves establish a theory of mind.

**Major Works**

- "An Unsolvable Problem of Elementary Number Theory" (1936) — effective calculability, lambda-definability, and an unsolvable problem.
- "A Note on the Entscheidungsproblem" (1936) — the absence of a general decision procedure for first-order validity.

- "A Formulation of the Simple Theory of Types" (1940) — a typed logical calculus.

**Lambda Calculus**

The lambda calculus represents abstraction and application of functions through a small formal vocabulary. The expression `λx.x` is the identity function; applying it to an expression returns that expression. More complex expressions can encode numbers and operations, and reduction rules explain how applications are evaluated.

The untyped calculus does not guarantee that every expression reaches a final result. Nonterminating reductions are possible. Its computational importance lies in the functions it can represent and compute, not a promise that all computations finish. Typed versions place constraints on admissible expressions and serve additional logical purposes.

Lambda-definable functions, general recursive functions, and Turing-computable functions coincide in the relevant mathematical sense. Equivalence between these formal systems can be proved. It supplies evidence for the claim that they capture effective computation, but that further claim connects formal definitions with an initially informal notion.

**The Church–Turing Thesis**

The **Church–Turing thesis** identifies effectively calculable functions with the functions computable by a Turing machine, equivalently by the other standard formalisms. "Effective" concerns a finite, precisely specified, step-by-step method executable without ingenuity at each step. Its original analysis concerns idealized calculation, not every possible process that can occur in nature.

The thesis is not just the proved equivalence of lambda calculus and Turing machines. The informal notion of an effective procedure is not itself one of the formal systems between which equivalence is demonstrated. Arguments for the thesis include the convergence of independently motivated formalisms and analyses of what mechanical calculation requires.

Descriptions of it as either "merely a conjecture with no support" or "a theorem proving all physical processes computable" miss this distinction. It has extensive mathematical and conceptual support within its intended domain. Broader claims require further assumptions.

**Church's Theorem and the Entscheidungsproblem**

Hilbert's **Entscheidungsproblem** asks for a general effective procedure deciding logical validity. Church showed that there is no such procedure for arbitrary first-order formulas. Turing's independent work arrived at a related negative result through a different formal analysis.

Undecidability does not mean that no first-order statements can be decided, that proofs cannot be checked, or that first-order logic lacks useful automated procedures. It means there is no algorithm that always finishes and correctly decides validity for every input formula. Validities can still be effectively enumerated, and restricted fragments can be decidable.

Nor is undecidability the same as computational expense. An algorithm can exist but be infeasible for large inputs; an undecidable problem admits no total decision algorithm of the required kind. Church's result concerns the latter boundary.

### Types and Logical Structure

Church's 1940 simple theory of types combines a hierarchy of types with lambda abstraction. Types distinguish kinds of expressions and restrict their combinations, avoiding unrestricted self-application and other constructions that create problems in untyped logical systems. This contribution matters to higher-order logic and formal reasoning as well as programming-language theory.

The typed logical project should not be conflated with the untyped calculus used in characterizing general computability. Related notation can serve different theoretical purposes.

### Later Strengthenings and Disputes

A **physical Church–Turing thesis** concerns what physically realizable devices can compute. The **extended Church–Turing thesis** concerns efficient simulation and computational complexity. A Church–Turing–Deutsch principle concerns universal computation in a physical theory. Authors formulate these proposals differently, and the labels "strong" and "physical" are not uniform.

Quantum computation bears directly on efficiency claims and the physical modeling of computation. Standard quantum computation does not straightforwardly refute the original thesis about effectively calculable functions. Proposals for hypercomputation instead examine whether some nonstandard physical resources could exceed Turing computability; their physical availability is disputed.

### Computation and Mind

A theory identifying mental activity with computation needs an account of which physical processes implement a computation, what representational content they carry, and which organization suffices for cognition or consciousness. Church's formal results do not answer those questions. They also do not show that artificial systems cannot have minds.

The distinction between simulating and realizing a process likewise needs substantive criteria. A mathematical description's computability does not automatically settle which properties its implementation shares with the described system. These are philosophical and scientific extensions of computability theory, not conclusions contained in Church's theorem.

### Further Reading


- Church, Alonzo. "An Unsolvable Problem of Elementary Number Theory." *American Journal of Mathematics* 58 (1936): 345–363.
- Church, Alonzo. "A Note on the Entscheidungsproblem." *The Journal of Symbolic Logic* 1 (1936): 40–41.
- Church, Alonzo. "A Formulation of the Simple Theory of Types." *The Journal of Symbolic Logic* 5 (1940): 56–68.
- Copeland, B. Jack. "The Church–Turing Thesis." *Stanford Encyclopedia of Philosophy*. Substantive revision 2023.

**Related Writing on This Site**

- *Mind, Matter, and Meaning* — Gordon Swobe's book-length argument about mind, consciousness, and meaning; an author position, not an endorsement by this reference page or its subjects.

**My View**

Church established a robust boundary for effective computation. Nothing in that result says that minds are abstract programs, that program identity establishes mental identity, or that computing a model of a process reproduces every causal power of the process modeled. The overreach belongs to later attempts to derive computationalism or artificial consciousness from a thesis about calculable functions.

The convergence of lambda calculus and Turing computation explains why abstract computability is substrate-indifferent. It does not follow that every mental capacity is individuated at that abstract level. A running implementation may realize content-fixing history, producer-consumer use, world-involving relations, memory, correction, planning, and control that its program type leaves open. Meaning depends on the relevant history and use; understanding on integration; consciousness on further mode-sensitive and poised phenomenal organization. Practical stake bears separately on autonomy and welfare. The book therefore rejects the inference from computability to mentality, not computation or artificial minds in principle.

**Tim Crane**

Tim Crane is a British philosopher whose work concerns intentionality, consciousness, and thought about things that do not exist. He defends an intentional understanding of mental life while distinguishing what a state presents from the mode in which it presents it. His discussions of nonexistent objects ask how ordinary thought and language can concern something without committing us to a corresponding entity.

**Major Works**

- *The Mechanical Mind* (first published 1995) — representation, computation, and theories of mind; substantially revised in later editions.
- "Intentionality as the Mark of the Mental" (1998) — a defense of intentionality as characteristic of mentality.
- *Elements of Mind: An Introduction to the Philosophy of Mind* (2001) — intentionality, consciousness, mental causation, and the mind–body problem.
- "The Intentional Structure of Consciousness" (2003) — intentional content, mode, and bodily sensation.
- *The Objects of Thought* (2013) — intentional objects and nonexistence.

**Intentionality and Aspectual Shape**

An intentional state concerns something under some aspect. A person can think of a city as a destination without thinking of it as the birthplace of a writer, even when both descriptions concern the same place. Specifying what the thought concerns is therefore not always sufficient to specify how it is presented.

Crane uses the distinction among **subject**, **intentional mode**, and **intentional content**. The subject bears the state. The mode distinguishes, for example, believing from hoping, or seeing from imagining. The content concerns what is presented and how things are represented. A subject can entertain the same content under different modes or different contents under one mode.

Intentionality in this broad sense is not restricted to propositional attitudes stated with "that" clauses. A bodily sensation or a perceptual presentation can have intentional structure even if its content is not naturally expressed as a complete proposition. This flexibility is central to Crane's treatment of consciousness.

### The Intentional Structure of Consciousness

Crane challenges a strict division between intentional states, such as belief, and purely qualitative states, such as pain. He argues that bodily sensation exhibits directedness: pain presents a part or region of the body under a distinctive mode. A felt pain in the foot illustrates the difference between the state and what it concerns. In phantom pain, the apparent bodily object may be absent without the sensation ceasing to occur.

This is an intentionalist account, but not one that identifies every phenomenal difference with a difference in represented worldly properties alone. Seeing a scene and imagining it can differ even when their contents overlap. On Crane's approach, mode belongs to the explanation of what the experience is like.

The strength of an intentionalist claim also matters. Saying that complete intentional structure determines phenomenal character does not automatically settle the metaphysical identity of properties or provide a reductive physical explanation.

### Transparency and Its Limits

Transparency arguments emphasize that attention to perceptual experience tends to disclose what it presents, such as a tree's color and shape, rather than an inner mental object. Crane's account preserves that directedness while asking whether introspection can also reveal the mode of presentation.

A difference between seeing and hearing is not necessarily a difference between the properties represented by those senses. Similarly, painful presentation differs from neutrally thinking about the same bodily region. These comparisons put pressure on content-only explanations without establishing an additional nonintentional "mental paint." The alternatives include richer contents, mode-sensitive intentionalism, and theories allowing irreducible qualitative properties.

### Thinking of Nonexistent Objects

We can think of Sherlock Holmes, a golden mountain, or an impossible figure. These thoughts are not automatically unintelligible because their apparent objects do not exist. Yet treating intentionality as an ordinary relation between two existing things appears to leave such cases unexplained.

Crane resists introducing a special domain of nonexistent entities simply to provide objects for every thought. Calling something an **intentional object** identifies what a state is about; it need not assign that object membership in a special ontological category. A person thinking of Holmes is not thereby related to a real detective.

This approach requires distinctions among the object of thought, the content of the thought, and whatever existing things help explain it—such as a book, a description, or a mental representation. None of those existing things automatically replaces Holmes as what the thought concerns. The theory aims to respect ordinary discourse while remaining cautious about existential commitments.

### Criticisms and Replies

Critics of intentionalism ask whether moods and sensations really present objects or whether intentional language is being imposed on them. Crane's reply broadens intentionality beyond propositional attitudes and uses bodily and aspectual structure as evidence. A further objection asks whether mode provides an explanation of qualitative differences or merely renames them. The account then needs an independent analysis of the relevant modes and their connection to experience.

Object theorists question whether a theory avoiding nonexistent entities can explain apparently true statements about them. Other critics favor reducing such discourse to descriptions or fictional contexts. Crane's project distinguishes the truth of relevant claims from an assumption that all apparent objects must exist, while acknowledging that different kinds of nonexistent-object discourse require analysis.

### Development and Scope

The 1998 defense of the mental mark leads into the 2003 analysis of conscious intentional structure; the 2013 book develops the problem of objects more extensively. These projects connect, but an account of aboutness is not automatically a full account of consciousness, computation, or mental causation. Crane's work makes those connections available for argument rather than treating them as conclusions of terminology.

### Related Concepts and Figures

- *Intentionality* — content, object, and mode.
- *Intentionalism* — intentional structure and phenomenal character.
- *Qualia* — qualitative properties and their explanation.
- *Transparency* — introspection of experience.
- *Franz Brentano* — the historical mark-of-the-mental thesis.
- *Ned Block* — objections to representational accounts.

### Further Reading


- Crane, Tim. *The Mechanical Mind: A Philosophical Introduction to Minds, Machines and Mental Representation*. First edition. Penguin Books, 1995.
- Crane, Tim. "Intentionality as the Mark of the Mental." In *Contemporary Issues in the Philosophy of Mind*, edited by Anthony O'Hear. Cambridge University Press, 1998.

- Crane, Tim. *Elements of Mind: An Introduction to the Philosophy of Mind*. Oxford University Press, 2001.
- Crane, Tim. "The Intentional Structure of Consciousness." In *Consciousness: New Philosophical Perspectives*, edited by Quentin Smith and Aleksandar Jokić. Oxford University Press, 2003.
- Crane, Tim. *The Objects of Thought*. Oxford University Press, 2013.

**Related Writing on This Site**

- *Mind, Matter, and Meaning* — Gordon Swobe's book-length argument about mind, consciousness, and meaning; an author position, not an endorsement by this reference page or its subjects.

**My View**

Crane is a close interlocutor because his intentionalism preserves the world-directed character of consciousness without adding non-intentional mental paint. His correction to strong transparency is one the book now accepts: introspection can disclose not only what a state presents but the intentional mode under which it presents it — seeing, hearing, fearing, hurting, imagining.

The remaining difference concerns identity. Crane says phenomenal character is determined by intentional mode and intentional content without making the book's explicit numerical-identity claim. I go further: within a state already counted as conscious and borne by an identified subject, phenomenal character consists in the complete world-presenting intentional profile — independently fixed content under a perceptual or bodily-affective mode, with determinacy, field structure, and attentional organization. That is a substantive extension of Crane's position, not the correction of an unmotivated hedge.

On the objects of thought: Crane's treatment of nonexistent objects is subtle and largely persuasive. The book doesn't dwell on this problem, focused as it is on veridical and near-veridical perception rather than thought about fictional objects, but Crane's framework is the right one for the cases outside that focus. See *Intentionality*.

**Francis Crick and Christof Koch**

Francis Crick (1916–2004) and Christof Koch are scientists whose collaboration helped develop an experimental program for studying the neural basis of consciousness. Crick brought a background in molecular biology to neuroscience; Koch's work concerned neural computation and visual processing. Their joint approach emphasized testable hypotheses about conscious vision rather than waiting for agreement on every philosophical question about mind.

**Major Works**

- Crick and Koch, "Towards a Neurobiological Theory of Consciousness" (1990) — visual awareness, attention, binding, and neural oscillations.
- Crick, *The Astonishing Hypothesis: The Scientific Search for the Soul* (1994) — a reductionist approach to mind and brain.
- Crick and Koch, "A Framework for Consciousness" (2003) — an organized set of proposals about neural mechanisms of visual consciousness.
- Koch, *The Feeling of Life Itself: Why Consciousness Is Widespread but Can't Be Computed* (2019) — an account informed by integrated information theory.

These works span joint and individual projects. Koch's later commitments should not be retroactively attributed to Crick or treated as the shared position of their early collaboration.

### Neural Correlates of Consciousness

A **neural correlate of consciousness**, or NCC, is commonly characterized as a minimal neural mechanism sufficient for a particular conscious experience under suitable background conditions. Researchers distinguish a correlate of a specific content—such as seeing a face—from the conditions that enable consciousness generally.

The strategy compares cases where a subject's experience changes while some aspects of the stimulus are controlled. In binocular rivalry, for instance, different images supplied to the two eyes can produce alternating perceptual dominance while the visual input remains relatively stable. Neural activity that tracks the percept rather than just the input provides a candidate correlate.

A correlation is not yet a complete causal explanation. Activity may enable the experience, constitute a relevant mechanism, or result from decisions, attention, memory, and reporting. Experimental interventions, changes in task, and converging measurements help distinguish these possibilities. The term "NCC" identifies a research target, not a certificate that any correlated brain signal is consciousness itself.

### The 1990 Proposal: Binding and Oscillations

Crick and Koch focus on visual awareness as a tractable starting point. They distinguish a fast form associated with iconic memory from a slower form involving attention and short-term memory. For the slower form, they propose that an attentional mechanism binds neurons representing features of one object through coherent, partly synchronized oscillations, probably in the 40–70 Hz range.

This was a hypothesis linking perceptual organization with neural dynamics, not an established universal frequency of consciousness. Gamma-range activity also appears in processes not uniquely associated with conscious experience. The importance of the proposal lies partly in its experimental specificity: it offered physiological mechanisms and tests rather than declaring that a difficult phenomenon lay outside scientific inquiry.

### The 2003 Framework

"A Framework for Consciousness" organizes proposals about conscious vision in terms of competing cellular assemblies. The aim is to connect neural representations with which content becomes consciously experienced, while distinguishing an explanatory framework from a finished theory.

The framework's scope is deliberately limited. A proposal for visual awareness does not immediately explain every modality, self-consciousness, or the complete unity of a subject's experience. The paper acknowledges aspects of the brain left outside the scheme. Presenting it as a final discovery of one consciousness center would misdescribe both its method and its claims.

### Crick's Reductionism

In *The Astonishing Hypothesis*, Crick argues that mental life can be understood through the activity of nervous tissue and its organization. This is a reductionist scientific program. It should not be equated without further argument with denying that experiences occur: explaining an activity in terms of mechanisms differs from eliminating the activity as unreal.

Critics question whether identifying mechanisms explains why their activity has a particular felt character. Crick treats neural correlates as a necessary starting point and considers the difficulty of communicating qualitative experience. A philosophical explanatory gap and an experimentally incomplete mechanism are distinguishable problems, even when both motivate further inquiry.

### Koch's Later Integrated-Information Approach

In his 2019 book, Koch argues that correlates need a theory relating physical organization to subjective experience. He draws on Giulio Tononi's **integrated information theory** (IIT), which begins with features of experience and proposes corresponding intrinsic cause–effect organization in physical systems.

On this approach, consciousness depends on a system's actual causal powers, not simply on which input–output computation it implements. Koch consequently distinguishes a software simulation of a brain from a physical system with the relevant causal organization. This is a claim within the theory, not an independently established verdict that all computers are unconscious.

IIT can attribute experience much more widely than human brains, but it does not simply say that every object or aggregate is conscious. Its criteria concern integration and the selection of an appropriate causal whole. Critics question the link from the theory's phenomenological starting points to its physical formalism, its counterintuitive attributions, and the feasibility of measuring the relevant quantities in complex systems. Defenders regard the formal commitments as a source of explanatory and empirical constraints.

### What the Program Leaves Open

Finding neural mechanisms can constrain theories of perception and consciousness without resolving every metaphysical dispute. Physicalists, dualists, and other theorists can agree about a correlate while disagreeing about what it explains. Conversely, philosophical disagreement does not make all empirical theories equally supported. Their predictions about particular mechanisms need testing on their own terms.

The history is therefore not a forced choice between Crick's reductionism and Koch's later theory. Workspace, recurrent-processing, higher-order, and other approaches formulate distinct hypotheses. The joint NCC program and the later IIT project contribute different kinds of explanation and require different evidence.

### Related Concepts and Figures

- *The Hard Problem of Consciousness* — neural mechanisms and phenomenal explanation.
- *Identity Theory* — identification of mental and physical states.
- *Physicalism* — reduction and its alternatives.
- *Panpsychism* — comparisons with widespread-consciousness proposals.

- *The Explanatory Gap* — what correlates alone establish.
- *Stanislas Dehaene* — a workspace-based research program.

**Further Reading**


- Crick, Francis, and Christof Koch. "Towards a Neurobiological Theory of Consciousness." *Seminars in the Neurosciences* 2 (1990): 263–275.
- Crick, Francis. *The Astonishing Hypothesis: The Scientific Search for the Soul.* Charles Scribner's Sons, 1994.
- Crick, Francis, and Christof Koch. "A Framework for Consciousness." *Nature Neuroscience* 6 (2003): 119–126. doi:10.1038/nn0203-119.
- Koch, Christof. *The Feeling of Life Itself: Why Consciousness Is Widespread but Can't Be Computed.* MIT Press, 2019.


**Related Writing on This Site**

- *Mind, Matter, and Meaning* — Gordon Swobe's book-length argument about mind, consciousness, and meaning; an author position, not an endorsement by this reference page or its subjects.

**My View**

Crick and Koch did the field an enormous service by making consciousness a laboratory problem, and the book inherits their naturalism without embarrassment. With the NCC program as method I have no quarrel at all — finding which neural events go with which experiences is exactly the right empirical work. The fork comes at the gap the correlates leave open: Crick bet that reduction would simply arrive; Koch concluded it would not, and reached for a theory that makes experience fundamental. The book takes a third road. I am closer to Crick than to late Koch — but not all the way to Crick. His astonishing hypothesis is physicalism with an eliminativist accent, the "nothing but a pack of neurons" framing that the book refuses. World-presenting experience is preserved on my account, not unmasked as an illusion of the cells; the reduction is real but *non*-eliminative. So I take Crick's one world and drop his "nothing but."

Koch's journey is the more instructive cautionary tale, because he followed the evidence where it embarrassed him. He looked hard at the correlates, saw that they never deliver *why* there is something it is like, and concluded the felt quality must be a fundamental feature of integrated matter. That is the panpsychist slide the book diagnoses — the explanatory gap, mistaken for a hole in the world and then filled with a new fundamental ingredient. His lost bet to Chalmers is the program's honesty made public: the correlates did not close the gap, and they were never going to, because the gap is not a missing mechanism. It is epistemic — a feature of how we conceive a single physical fact under two modes of access. Find every correlate and the puzzle remains, not because something physical is missing but because the first-person concept and the third-person concept never visibly coincide. Crick under-read the gap; Koch over-read it. The book reads it as the shape of a concept, and deflates it.

**Stanislas Dehaene**

Stanislas Dehaene is a French cognitive neuroscientist whose research concerns numerical cognition, reading, and consciousness. With collaborators including Jean-Pierre Changeux, Michel

Kerszberg, and Lionel Naccache, he developed the **global neuronal workspace** (GNW) approach to conscious processing. It connects a functional idea about shared information with hypotheses about recurrent activity and long-range connections in the brain.

**Major Works**

- *The Number Sense* (1997) — numerical cognition and its neural basis.
- *Reading in the Brain* (2009) — reading and the adaptation of existing neural systems to literacy.
- *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts* (2014) — conscious access and the workspace framework.
- Dehaene, Kerszberg, and Changeux, "A Neuronal Model of a Global Workspace in Effortful Cognitive Tasks" (1998) — a neural architecture and simulation of effortful processing.
- Dehaene and Naccache, "Towards a Cognitive Neuroscience of Consciousness" (2001) — empirical evidence and a workspace framework.
- Mashour, Roelfsema, Changeux, and Dehaene, "Conscious Processing and the Global Neuronal Workspace Hypothesis" (2020) — a review of the theory and its evidence.

The 1998 model was coauthored by Kerszberg, not Naccache. The papers represent stages of a collaborative research program rather than one result discovered by a single investigator.

**The Workspace Proposal**

GNW distinguishes specialized processors from a distributed network capable of sharing their information. Many perceptual and cognitive operations can proceed locally without conscious access. When a representation is sufficiently amplified and sustained through recurrent processing, it becomes available to otherwise separate systems for memory, evaluation, reasoning, flexible action, and report.

The workspace is not a single anatomical room or an inner observer. In the 1998 model, neurons with long-range connections link distributed processors, with particularly substantial contributions proposed for prefrontal and inferior parietal regions. Descending connections can enhance or suppress selected processing. The resulting architecture aims to explain flexible coordination rather than introduce a place where all information is displayed to a spectator.

The functional **global workspace** idea has a history including Bernard Baars's work. Dehaene and collaborators' specifically neuronal proposal adds assumptions about anatomy, dynamics, and measurable neural activity. Related workspace theories need not accept every GNW anatomical commitment.

**From Effortful Tasks to Conscious Access**

The 1998 paper models the Stroop task, in which naming a word's ink color can conflict with reading the word itself. It predicts increased workspace recruitment during learning, effortful performance, and after errors. This connects the architecture to cognitive control; it is not by itself an experiment establishing the location of every conscious experience.

Later GNW work contrasts conscious and nonconscious processing using masking, attentional limitations, and other methods. A briefly presented stimulus can influence later behavior even

when a person does not report seeing it. Conscious access is associated, in the framework, with a transition from such local processing to amplified, sustained, broadly available representation.

**Ignition** names the proposed nonlinear transition. Recurrent activity permits a representation to persist and affect distant processors instead of disappearing with the initial sensory response. In many experiments, late widespread activity occurs on the scale of a few hundred milliseconds. The timing and measured signature depend on the task, stimulus, and measurement; "300 milliseconds" is not a universal law or a literal clock marking the onset of all experience.

### Evidence and Methodological Problems

GNW draws on behavioral contrasts, neuroimaging, electrophysiology, and changes between conscious and unconscious states, including anesthesia. These methods address different questions: whether a particular content is consciously accessed, whether a person is in a state supporting consciousness, and which processes permit later use of information.

A central difficulty is distinguishing consciousness from the processes needed to report it. Asking someone to indicate that they saw a stimulus recruits decision, memory, action preparation, and monitoring. Some late activity may reflect those additional demands. No-report approaches seek alternative indicators, but their validity also requires justification: a proxy is not automatically a direct measure of experience.

GNW proponents distinguish global availability from an actual spoken report and use multiple forms of evidence to support the mechanism. Critics dispute whether prefrontal activity and particular late signals are necessary for every conscious percept, or whether recurrent activity in more restricted sensory networks can suffice. These questions should not be settled by treating any single neural signature as identical with consciousness.

### Phenomenal Consciousness and Access

GNW is intended as an account of conscious processing, not merely an account of speech. Nevertheless, philosophers distinguish the mechanisms of access from phenomenal character—what an experience feels like. Ned Block's overflow arguments question whether all experienced detail must be globally available to limited-capacity cognition.

Workspace defenders question whether the evidence shows conscious but inaccessible detail rather than unconscious information that becomes available after a cue. More broadly, they argue that access, flexible use, and neural dynamics provide the appropriate scientific framework for explaining consciousness. Critics ask whether that framework captures experience itself or primarily its cognitive consequences.

It would therefore be misleading either to announce that broadcasting has conclusively explained all phenomenology or to define GNW as a theory its proponents concede has nothing to say about experience. The disagreement concerns whether, and in what form, workspace organization is necessary or sufficient for phenomenal consciousness.

### Other Theories and Development

Integrated information theory begins from a different proposed relation between experience and intrinsic causal organization. Recurrent-processing theories emphasize feedback within sensory processing; higher-order theories emphasize representations of one's mental states. Overlap in vocabulary does not remove differences in predictions about the processes and regions required.

GNW has developed from an architecture for effortful processing into a broader program addressing conscious contents, conscious states, and clinical assessment. Continued testing can revise particular signatures or anatomical proposals without automatically confirming or refuting every global-workspace theory. Claims of a final winner require specifying which version, prediction, and experiment is at issue.

### Related Concepts and Figures

- *Functionalism* — explanation through availability and cognitive roles.
- *The Hard Problem of Consciousness* — phenomenal and functional explanatory tasks.
- *Higher Order Thought Theories* — an alternative account of conscious states.
- *The Explanatory Gap* — neural explanations and felt character.
- *Ned Block* — access, phenomenality, and overflow.
- *Francis Crick and Christof Koch* — neural correlates and competing programs.

### Further Reading


- Dehaene, Stanislas. *The Number Sense: How the Mind Creates Mathematics*. Oxford University Press, 1997.
- Dehaene, Stanislas. *Reading in the Brain: The Science and Evolution of a Human Invention*. Viking, 2009.
- Dehaene, Stanislas. *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. Viking, 2014.
- Dehaene, Stanislas, Michel Kerszberg, and Jean-Pierre Changeux. "A Neuronal Model of a Global Workspace in Effortful Cognitive Tasks." *Proceedings of the National Academy of Sciences* 95 (1998): 14529–14534.
- Dehaene, Stanislas, and Lionel Naccache. "Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework." *Cognition* 79 (2001): 1–37.
- Mashour, George A., Pieter Roelfsema, Jean-Pierre Changeux, and Stanislas Dehaene. "Conscious Processing and the Global Neuronal Workspace Hypothesis." *Neuron* 105 (2020): 776–798.


### Related Writing on This Site

- *Mind, Matter, and Meaning* — Gordon Swobe's book-length argument about mind, consciousness, and meaning; an author position, not an endorsement by this reference page or its subjects.

### My View

Global broadcasting gives a powerful account of access: how information becomes available for report, memory, reasoning, and control. The global-ignition evidence is real and important. What that evidence directly establishes is a candidate mechanism for conscious access, not an identity

between broadcasting and phenomenal character. Whatever broader philosophical ambitions one attaches to GNW, the book keeps those claims separate.

The book and Dehaene can therefore mostly share a roof. I hold that the felt character of experience consists in world-directed representational content of the right kind, and that the residual mystery is epistemic, not a further mechanism waiting to be found. Global broadcasting is plausibly part of the machinery by which such content gets poised for use — Tye's "poised" condition has a natural home in a workspace. So I read Dehaene not as a competitor to the identity claim but as supplying some of its neural detail, provided we do not mistake the broadcasting *mechanism* for the felt character itself. The one place I dig in is against the slide some of his readers make — from "consciousness is global availability" to "and that is all consciousness *is*." Availability is what the brain does with a content; it is not the same as the content's being world-presenting. Keep those apart and Dehaene's science and the book's philosophy fit together better than either fits IIT.

**Daniel Dennett**

Daniel C. Dennett (1942–2024) was an American philosopher whose work connected philosophy of mind with cognitive science and evolutionary theory. He developed naturalistic accounts of consciousness, intentionality, free will, and the self. His recurring question was how organized physical processes could produce the capacities associated with minds without requiring a further inner observer.

**Major Works**

- *Content and Consciousness* (1969) develops an early approach to mental content and consciousness.
- *The Intentional Stance* (1987) collects arguments about belief, desire, interpretation, and prediction.
- "Quining Qualia" (1988) challenges traditional conceptions of experiential qualities.
- *Consciousness Explained* (1991) presents the Multiple Drafts model and heterophenomenology.
- "Real Patterns" (1991) explains his position on the reality of beliefs and other higher-level patterns.
- "The Self as a Center of Narrative Gravity" (1992) develops his account of personal selfhood.
- *From Bacteria to Bach and Back* (2017) extends his evolutionary account of cognition and cultural understanding.

**Consciousness Without a Cartesian Theater**

Dennett calls the supposed central place where everything becomes available to an inner witness the **Cartesian Theater**. Rejecting an immaterial soul does not by itself remove this picture: a materialist may still imagine a privileged brain location where experience is displayed and finally becomes conscious.

The **Multiple Drafts model** replaces that finish line with distributed, partly competing processes. Sensory information undergoes interpretation and revision in different places and at different times. What a person later reports depends on which processes influence memory, judgment, and action. Dennett argues that there need not be a further fact about which version appeared before

a spectator at an exact instant. The model challenges the assumption that consciousness must have a single, determinate presentation sequence independent of its subsequent effects.

His later work places these capacities in an evolutionary setting. Organisms can perform useful tasks without understanding why those tasks work: **competence without comprehension**. Evolution, learning, language, and culture can gradually build more reflective forms of understanding from such competences. This is an account of their development, not a claim that human understanding does not exist.

### Heterophenomenology

Heterophenomenology is Dennett's method for incorporating first-person testimony into a third-person science. Reports of pain, imagery, or an apparently continuous visual field are evidence to explain. The researcher provisionally constructs the world as the subject describes it without assuming that every reported feature accurately characterizes the underlying processes.

This separates taking testimony seriously from treating introspection as infallible. A subject may sincerely report a seamless experience even if the mechanisms producing that report are fragmentary. Dennett regards the method as a starting point for investigation, to be supplemented by psychology and neuroscience, rather than as the entire investigation.

### Intentional Stance and Real Patterns

The intentional stance predicts a system's behavior by treating it as a reasonably rational agent with beliefs and desires. A chess computer, for example, can be understood as trying to protect its king without containing a sentence-shaped belief for every prediction an observer makes.

Dennett does not therefore regard all belief attribution as arbitrary fiction. In "Real Patterns," a pattern can be objective when it supports economical description and reliable prediction. Beliefs may be real at this level even if neuroscience finds no isolated object corresponding to each belief. His position resists both the demand that beliefs be discrete inner items and the conclusion that higher-level patterns must be unreal.

### Qualia and the Narrative Self

"Quining Qualia" targets qualities conceived as intrinsic, ineffable, private, and directly apprehensible. Dennett's examples of changing tastes, uncertain memories, and altered reactions question whether an experiential quality can be isolated from judgments and dispositions in the manner this conception requires. He explicitly grants that conscious experiences have properties; his claim is that those properties do not satisfy the traditional description of qualia. His relation to later illusionism should be understood through this distinction rather than summarized as a denial that anyone experiences anything.

The self as a **center of narrative gravity** is an abstraction generated by an organism's narratives and interpretive organization. Like a center of gravity, it need not be a material part located somewhere inside the system. Dennett also compares selves with fictional protagonists. How this analogy relates to his realism about patterns is a substantive interpretive question: an abstraction's usefulness and objectivity need not amount to the existence of an inner owner.

### Criticisms and Replies

John Searle and other critics argue that explaining discrimination, report, and behavioral control leaves subjective experience unexplained. Dennett replies that the demand for something over and above these processes often imports the theater or the disputed traditional conception of qualia. The disagreement concerns what a successful explanation must preserve, not simply whether the brain matters.

Phenomenologists such as Dan Zahavi question whether heterophenomenology brackets the first-person perspective so thoroughly that it loses the phenomenon it seeks to study. Dennett's response is that scientific neutrality does not require ignoring testimony; it requires investigating the mechanisms and possible errors behind it. Critics remain unconvinced that this exhausts the evidential contribution of lived experience.

Accounts of minimal selfhood also question whether narrative explains every form of subjectivity. Infants and nonlinguistic animals pose a particular challenge to making self-narration a general condition of having a perspective. Dennett's narrative account principally addresses the developed personal self; its scope must be distinguished from the broader question of consciousness in animals.

### Related Concepts and Figures

- *The Inner Theater* — the central observer model Dennett criticizes.
- *Functionalism* — explaining mental states through their organization and effects.
- *Qualia* — disputes over the properties of conscious experience.
- *The Self* — narrative, bodily, and other accounts of selfhood.
- *John Searle* — a critic of Dennett's explanation of consciousness.
- *Ned Block* — the distinction between phenomenal consciousness and cognitive access.

### Further Reading


- Dennett, Daniel C. *Consciousness Explained*. Little, Brown, 1991.
- Dennett, Daniel C. "Quining Qualia." In *Consciousness in Modern Science*, edited by A. J. Marcel and E. Bisiach. Oxford University Press, 1988.
- Dennett, Daniel C. "Real Patterns." *The Journal of Philosophy* 88, no. 1 (1991): 27–51.
- Dennett, Daniel C. "The Self as a Center of Narrative Gravity." In *Self and Consciousness: Multiple Perspectives*, edited by F. Kessel, P. Cole, and D. Johnson. Erlbaum, 1992.


### Related Writing on This Site

Mind, Matter, and Meaning presents the site author's arguments about mind and meaning. It is a separate argumentative work, not a neutral source for the account above.

### My View

Dennett is my closest ally on the diagnostic side and one of my sharpest opponents on the constructive side. No one did more to discredit the Cartesian Theater, and the book's demolition runs beside his. His center of narrative gravity supplies the nearest rival to the assembled self: I accept that the self is constructed through narrative and organization, but reject the verdict of

fiction. A pattern integrated across memory, correction, planning, commitment, and action does causal work in the life it reports and qualifies as real by Dennett's own real-patterns standard.

The break comes when Dennett follows his demolition of the theater all the way to *illusionism*: the view that phenomenal character, as ordinarily conceived, is a kind of user-illusion to be explained away rather than relocated. There the argument stops and turns the other way. Clearing the inner theater does not abolish phenomenal character; it re-sites it as the representational content of world-directed experience. The phenomenology is real. Only its inner-object reading goes.

Crane and Farkas put the instinct well: "where consciousness is concerned, we depart from Dennett's interpretationism." I take Dennett's wrecking ball and decline his replacement. Walk with him to tear down the theater; part from him over whether anything real is left standing inside experience.

**René Descartes**

René Descartes (1596–1650) was a French philosopher and mathematician whose work joined a new mechanistic science with a search for secure foundations of knowledge. His method of doubt, account of a thinking subject, and distinction between mind and body shaped later debates about consciousness and the external world. Earlier traditions had already investigated soul, thought, and perception; Descartes reorganized those questions rather than originating them all.

**Major Works**

- *Discourse on the Method* (1637) introduces his method and appeared with scientific essays, including the *Geometry*.
- *Meditations on First Philosophy* (1641), with Objections and Replies, develops the doubt, the cogito, arguments for God, and the real distinction between mind and body.
- *Principles of Philosophy* (1644) presents a systematic account of metaphysics and natural philosophy.
- *The Passions of the Soul* (1649) examines emotions and embodied human life.
- His correspondence with Princess Elisabeth of Bohemia, beginning in 1643, examines mind–body interaction and union.

**Doubt and the Cogito**

The doubt in the *Meditations* is methodological. Descartes provisionally sets aside beliefs that admit doubt to discover whether anything survives. Sensory error calls particular judgments into question; dreaming challenges confidence in one's present surroundings; the possibility of a powerful deceiver extends doubt to apparently evident reasoning.

The act of doubting nevertheless establishes the existence of the thinker while that thought occurs. Meditation II formulates this as the certainty that "I am, I exist" whenever thought or asserted. The familiar *cogito ergo sum* appears in other formulations of the argument. Commentators disagree about whether the cogito is an inference, an immediate intellectual recognition, or a performative certainty. It initially establishes neither the existence of a body nor a full account of the enduring person.

### Wax, Ideas, and Knowledge

A piece of wax changes smell, color, hardness, and shape when heated, yet is judged to be the same wax. Descartes argues that the intellect grasps its capacity for indefinitely many changes more adequately than a particular sensory image can. The example supports his claim that knowledge of body involves intellectual judgment and that the mind is better known than body.

This epistemic priority should not be equated without argument with a literal inner screen. Descartes distinguishes ideas, their representational significance, and their causes; later interpretations of the "Cartesian" picture do not settle every issue about his own theory of perception.

Clear and distinct perception supplies the proposed standard for knowledge. Descartes argues that a nondeceiving God secures the reliability of properly exercised cognition. He also explains error through the will's extending judgment beyond what the intellect clearly understands. The project therefore combines an account of rational insight with an account of how a finite thinker can go wrong.

### Real Distinction and Human Union

In Meditation VI, mind can be clearly and distinctly conceived as thinking and unextended, and body as extended and nonthinking. Descartes argues that God can create each apart from the other, establishing their **real distinction**. He also argues that mind is indivisible in a way an extended body is not. These are metaphysical arguments, not merely the observation that mental and bodily vocabulary differ.

Yet Descartes does not describe ordinary human life as a detached intellect piloting a machine from outside. Hunger, thirst, and pain indicate a close union: the thinker is not related to the body merely as a sailor to a ship. The senses have a practical role in this embodied life even when they do not reveal the fundamental nature of matter.

His physiological writings give bodily processes mechanistic explanations. In *The Passions*, the pineal gland has a special role in coordinating bodily and mental activity. Identifying a bodily site does not, however, explain how an unextended mind causally interacts with extended matter.

### Objections and Descartes's Responses

**The Cartesian circle.** Arnauld asks whether the proof of God's reliability already depends on trusting clear and distinct reasoning, whose trustworthiness the proof is supposed to establish. Descartes distinguishes present, compelling intellectual insight from reliance on remembered demonstrations when attention has moved elsewhere. Whether this distinction resolves the circle remains disputed.

**Interaction.** Elisabeth asks how something without extension can determine bodily motion. Descartes responds that mind, body, and their union involve different primitive notions, and that the union is familiar in ordinary embodied experience. Critics argue that familiarity with interaction does not supply a satisfactory causal explanation. Defenders emphasize that his theory deliberately recognizes union as something not reducible to either substance considered alone.

**Conceivability and possibility.** Critics question whether apparently clear separation establishes genuine metaphysical separability. A thinker might understand something incompletely

and mistake that incompleteness for independence. Descartes's reply depends on the adequacy of clear and distinct understanding and the divine guarantee; rejecting those premises weakens the argument without making it a simple verbal mistake.

**Animals.** His denial of animal thought and his mechanistic explanations of animal behavior provoked objections about the evidence for mentality. The exact implications for animal sensation and pain require care: attributing his view as a blanket denial of every physiological form of sensation oversimplifies the texts and their interpretation.

### Related Concepts and Figures

- *Dualism* — the distinction between thinking and extended substances.
- *The Inner Theater* — a later model associated with the Cartesian inheritance.
- *Indirect Realism and Sense Data* — subsequent debates about ideas and perceptual mediation.
- *John Locke* — an empiricist response to rationalist theories of knowledge.
- *David Hume* — a challenge to knowledge of an enduring substantial self.
- *Immanuel Kant* — a later reconstruction of the conditions of objective knowledge.

### Further Reading

- Descartes, René. *Meditations on First Philosophy* (1641), including the Objections and Replies.
- Descartes, René. *Discourse on the Method* (1637).
- Descartes, René. *Principles of Philosophy* (1644).
- Descartes, René. *The Passions of the Soul* (1649).
- "Descartes' Epistemology." *Stanford Encyclopedia of Philosophy*.

### Related Writing on This Site

Mind, Matter, and Meaning presents the site author's arguments about mind and meaning. It is a separate argumentative work, not a neutral source for the account above.

### My View

Descartes gives the topic this book is about its distinctively modern architecture. Philosophers had investigated soul, intellect, perception, and thought for centuries before him; what changes with Descartes is the systematic placement of certainty inside a self-knowing, unextended mind and of the world outside it as something to be recovered. Modern philosophy of mind inherits that arrangement even where it rejects his conclusions.

What the book rejects, decisively, is the inheritance the *Meditations* leaves behind: substance dualism (because naturalism); the epistemic priority of mind over world (because transparency — when I attend to my experience of the wax, what I find is the wax, not an inner representation of it); and the inner theater that follows from the priority claim (the bad picture the whole book argues against). Dennett later named the picture the "Cartesian theater"; our culture inherited it from Descartes and has not yet shaken it off. The book's whole project is the slow archaeology of digging back out from under it.

What the book recruits, with credit: the method of doubt as a tool (even when the conclusion runs the other way); the seriousness with which Descartes took first-person experience as data; the willingness to follow an argument into uncomfortable places. The cogito itself stays a real insight

(the thinker, thinking, knows that thinking is happening), even when the priority claim piled on top of it is wrong. And the fact that Descartes gave the modern mind-body problem its canonical architecture means he set the table, even if everything he served has to be sent back.

The truest summary of the relationship is the one Ryle put first: the ghost in the machine is the picture to dissolve, not the picture to defend. But the ghost wouldn't even be visible to dissolve if Descartes hadn't put it there for everyone to see. See *The Inner Theater*, *Dualism*, and *Anti-Reification*.

**Fred Dretske**

Fred Dretske (1932–2013) was an American philosopher whose work moved from perception and knowledge to naturalistic accounts of representation, mental causation, and consciousness. His central project was to explain how a physical state can carry information, acquire a representational function, and sometimes misrepresent its surroundings. His informational approach should be distinguished both from bare correlation theories and from interpretations that make representation entirely dependent on an observer.

**Major Works**

- *Seeing and Knowing* (1969) examines perception and knowledge.
- *Knowledge and the Flow of Information* (1981) connects information with knowledge and mental content.
- "Misrepresentation" (1986) addresses how naturally occurring representations can be false.
- *Explaining Behavior: Reasons in a World of Causes* (1988) develops an account of the causal relevance of content.
- *Naturalizing the Mind* (1995) presents his representational approach to mental life and experience.

**Information and Representation**

In Dretske's technical account, a signal carries information that a condition obtains when, given the relevant channel conditions, the signal's occurrence guarantees that condition. This is stronger than a loose statistical association. Information can occur without anyone interpreting it: a natural indicator may convey information about an environmental condition even if no organism notices it.

Representation requires a further step. A system must have the **function of indicating** a property. Dretske's speedometer example separates this function from the instrument's actual success. A pointer position represents a particular speed because the instrument is supposed to indicate speed. When it malfunctions, it still represents that speed although the vehicle is moving differently.

The distinction makes error possible. Information, in the relevant factive sense, cannot be false; a state with an information-carrying function can fail to perform it. Applying this account to organisms requires explaining how such functions arise through biological history or learning, rather than simply importing a human designer's intentions into every case.

### Learning, Content, and Mental Causation

Dretske distinguishes functions belonging to a system's inherited organization from functions acquired through learning. In the latter case, a state is recruited to control behavior because it has carried useful information about a condition. Its subsequent occurrence can continue to represent that condition even when an unusual situation disrupts the original correlation.

This gives representational content an externalist dimension: what a state represents can depend on its environmental relations and recruitment history, not just its present physical structure. Two internally similar states might therefore differ in content if their relevant histories differ.

In *Explaining Behavior*, Dretske asks how content can help explain action when physical events already have physical causes. His answer distinguishes the immediate production of a movement from the explanation of how a behavior-controlling system acquired its organization. The information carried during learning helps explain why the system came to respond as it does. Critics question whether this historical role gives content enough causal relevance to present action; Dretske regards structuring explanations as genuine explanations rather than as competition with the immediate physical causes.

### Experience and the Representational Thesis

*Naturalizing the Mind* argues that mental facts are representational facts. For sensory experience, the qualities of which a subject is aware are properties represented in the environment or body. An experience of a red, round object presents redness and roundness; it need not contain an inner red, round object for the subject to inspect.

Dretske treats experience as nonconceptual representation. A creature can perceptually represent a property without possessing a concept with which to judge that property. This permits a distinction between sensory awareness and knowing or believing what one sees.

His view converges with Michael Tye's representational account, but their theories are not interchangeable. **PANIC**—Poised, Abstract, Nonconceptual, Intentional Content—is Tye's formulation, not Dretske's. Dretske's explanatory framework centers on indicator functions and the distinction between systemic and acquired representation.

### Criticisms and Replies

**Which property is represented?** A state may correlate with several connected conditions: an environmental property, a proximal stimulus, and a later behavioral opportunity. Critics ask what privileges one as its content. Dretske appeals to the state's indication function and the circumstances of its recruitment. Teleosemantic alternatives place different weight on the downstream use of a state, so these approaches should not be collapsed into one theory.

**History and newly formed duplicates.** Historical accounts face thought experiments involving a physical duplicate with no ordinary learning or evolutionary history. Critics argue that the duplicate could still think or feel. Dretske's externalism resists identifying present internal similarity with sameness of representational function. The disagreement tests whether history is constitutive of mentality or merely explains its usual production.

**Phenomenal character.** Inversion cases, blurry vision, and experiences allegedly lacking representational content challenge the claim that representation accounts for everything experiential. Representationalists reply by examining differences in content and kind of representation. The further question of why some representations are conscious remains distinct from the question of what they represent; the fact that a speedometer represents does not make it conscious on Dretske's account.

### Related Concepts and Figures

- *Information and Meaning* — information versus representational content.
- *Intentionalism* — representational accounts of phenomenal character.
- *Teleosemantics* — biological and learned functions in theories of content.
- *Transparency* — awareness of represented properties rather than inner colored objects.
- *Michael Tye* — a related but distinct theory of consciousness.
- *Ruth Millikan* — a different emphasis on the functions of representational systems.

### Further Reading


- Dretske, Fred. *Knowledge and the Flow of Information*. MIT Press, 1981.
- Dretske, Fred. *Explaining Behavior: Reasons in a World of Causes*. MIT Press, 1988.
- Dretske, Fred. *Naturalizing the Mind*. MIT Press, 1995.
- Dretske, Fred. "Misrepresentation." In *Belief: Form, Content, and Function*, edited by Radu J. Bogdan, 17–36. Oxford University Press, 1986.


### Related Writing on This Site

Mind, Matter, and Meaning presents the site author's arguments about mind and meaning. It is a separate argumentative work, not a neutral source for the account above.

### My View

Dretske supplies one major route into the book's central argument. His representational thesis converges with Tye's contemporary identity claim from the informational side; the book adopts the result as part of a broader cumulative case rather than on either figure's authority.

The informational framework has a particular virtue: it makes the externalism about content hard to avoid. If representation is a natural causal-informational relation between a state and the world, then what a state is about is fixed by what it tracks in the environment, not by anything enclosed in the skull. This is exactly the lever the book uses to dismantle the inner-theater picture — the idea that experience presents us with inner objects or properties rather than the world itself.

The limitation: lawful information alone underdetermines content. Dretske's own appeal to learning and recruited function points beyond bare covariance, and the book continues that route through Millikan's producer-consumer organization and Burge's anti-individualism. It then keeps a second question open: content in a mechanism does not by itself establish perception, understanding, or consciousness by the larger system. Dretske opens the door; the book separates the rooms.

See *Intentionalism*, *Teleosemantics*, and *Michael Tye*.

### Karl Friston

Karl Friston is a British neuroscientist whose work includes statistical methods for brain imaging, dynamic causal modeling, and the free-energy principle. His theoretical research connects perception, action, learning, and biological self-organization through probabilistic models. The framework has applications at several scales, but its mathematical scope and its interpretation as a theory of mind require separate assessment.

### Major Contributions

- Statistical parametric mapping and related methods analyze patterns in neuroimaging data.
- Dynamic causal modeling compares hypotheses about interactions among neural systems.
- "The free-energy principle: a unified brain theory?" (2010) presents connections between the principle and several theories of brain function.
- *A free energy principle for a particular physics* (2019) extends the framework to statistically individuated systems and self-organization.

### The Free-Energy Principle

An adaptive organism ordinarily occupies a restricted range of states: its temperature, chemical concentrations, and sensory exchanges cannot vary without limit if it is to persist. Friston proposes a mathematical description of such systems in terms of their tendency to remain within characteristic states.

Here **surprise** is a technical quantity: the negative logarithm of a state's probability under a specified model or distribution. It does not mean the conscious feeling of being startled. **Variational free energy** is an information-theoretic quantity that bounds surprise and can be evaluated using an approximate probability distribution over hidden causes. It should not be identified with ordinary thermodynamic free energy merely because the terms share a name.

A system cannot directly inspect every hidden cause of its sensory input. It can nevertheless change an approximate model or change its sensory exchanges in ways that reduce the free-energy bound. Under the framework's assumptions, this connects biological persistence with processes that can be described as inference. The assumptions, choice of variables, and timescale are part of the account; it is too strong to say that every surviving object simply and consciously pursues one universal objective.

### Perception, Learning, and Active Inference

Perception updates estimates of the causes of sensory input. Learning changes more persistent parameters of a generative model. Action changes what the organism samples and how it interacts with its surroundings. In predictive-coding formulations, prediction errors help drive these adjustments.

Consider looking for a partly hidden object. The organism can revise its estimate of the object's location or move its eyes and body to obtain more informative input. Action does not merely force the world to match any belief the organism happens to hold; it can seek information and bring about conditions that accord with prior preferences or expected trajectories.

**Active inference** develops this combined treatment of perception and action. Models of policy selection may trade off preferred outcomes and uncertainty reduction. Expected free energy used in evaluating future policies must be distinguished from variational free energy associated with current inference. Treating both as undifferentiated "prediction error" obscures that distinction.

### Markov Blankets and System Boundaries

A Markov blanket is a set of variables that renders other sets conditionally independent. In Friston's applications, sensory and active states mediate between internal and external states. Conditional on the blanket states, internal and external states are statistically independent in the proposed partition. This statistical relation should not be confused with a claim that their dynamics never causally influence one another.

This provides a statistical way to describe a boundary, not necessarily a literal membrane. The framework can be applied to cells, organisms, and larger or nested systems. Identifying a blanket does not alone show that the enclosed system is a conscious subject. Nor does every statistical partition establish an organism's biological or cognitive individuality.

In his 2019 formulation, Friston proposes that internal states can be interpreted as representing or making inferences about external states. Whether this interpretation amounts to a psychological explanation, a useful mathematical description, or something in between is part of the philosophical debate.

### Scope, Objections, and Responses

**Principle versus process theory.** A broad mathematical principle may constrain possible models without specifying the mechanism of any particular brain. Critics ask what observations could falsify the framework and whether its generality permits too many incompatible implementations. A response distinguishes the organizing principle from specific predictive-coding or active-inference models. Those models make additional commitments and can be assessed against data, even if the principle itself has a different explanatory status.

**The apparent inactivity problem.** If reducing surprise were the only objective, remaining in a predictable dark room might appear ideal. The framework's response is that an organism's expected or preferred states include the conditions for its continued life and activity; hunger and the expected consequences of inaction are not removed by darkening the room. This answer depends on how those expectations and preferences are specified rather than following from a bare instruction to avoid novelty.

**Inference and representation.** Describing a system mathematically as minimizing free energy does not automatically settle whether it literally has beliefs or meaningful representations. Defenders seek explanatory continuity between biological regulation and cognition. Critics demand further grounds for attributing psychological content, agency, or consciousness. Neither accepting the formalism nor rejecting an expansive interpretation decides these questions by itself.

The principle also does not imply that perception is awareness of an internal model rather than of the world. Computational descriptions of the mechanisms and philosophical descriptions of perceptual objects address different questions.

**Related Concepts and Figures**

- *Predictive Processing and the Interface* — generative models and prediction error in perception.
- *Embodiment* — regulation through bodily and environmental interaction.
- *Information and Meaning* — statistical inference and semantic interpretation.
- *The Hard Problem of Consciousness* — distinctions between cognitive explanation and phenomenal consciousness.

**Further Reading**

- Friston, Karl. "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience* 11 (2010): 127–138.
- Friston, Karl. *A free energy principle for a particular physics*. arXiv:1906.10184, 2019.

**Related Writing on This Site**

Mind, Matter, and Meaning presents the site author's arguments about mind and meaning. It is a separate argumentative work, not a neutral source for the account above.

**My View**

Friston is the rival who is also, in one crucial respect, a confederate. Strip the free energy principle back to its biological core and it says something the book affirms: a living system holds itself together against conditions that can dissolve it. That precarious organization can help explain autonomy, individuation, practical interest, and welfare. It does not by itself explain what any state represents. Original content still requires a history and producer-consumer organization that privilege one correctness condition over its rivals.

Where I part company is the same place I part from the predictive-processing tradition generally. The formalism describes the *machinery* of a self-maintaining system; it does not license the conclusion that what such a system is aware of is its own internal model. "Minimizing prediction error" is a story about vehicles — about the regulating, not the regulated. And the principle's imperial version, where free energy explains brains and rocks and economies alike, buys generality by draining the content out: a frame that fits everything tells you little about the specific, world-involving achievement that meaning is. So I keep the embodied core and decline the metaphysics. Friston has, I think, formalized a candidate basis of stake. He has not thereby supplied content or shown that the mind is sealed inside a model of the world.

**Gilbert Harman**

Gilbert Harman (1938–2021) was an American philosopher who worked on reasoning, epistemology, ethics, language, and mind. His 1990 paper "The Intrinsic Quality of Experience" helped shape contemporary debates about the transparency of perception. It defended psychophysical functionalism by challenging how several objections characterize the properties encountered in experience.

**Major Works**

- *Thought* (1973) develops views about thought, meaning, and psychological explanation.
- *Change in View* (1986) distinguishes rational changes of belief from the relations studied by deductive logic.

- "The Intrinsic Quality of Experience" (1990) examines direct awareness of experiential qualities, phenomenal knowledge, and spectrum inversion.
- *Reasoning, Meaning, and Mind* (1999) collects work across these areas.

**Eloise and the Transparency of Experience**

Harman's example concerns Eloise looking at a tree. When she attends to what she sees, the colors and shapes appear as properties of the tree and its surroundings. They do not appear as colors and shapes belonging to her act of seeing. Turning attention toward her experience does not reveal another colored item between her and the tree.

This observation motivates what later literature calls **transparency**. Reflection on visual experience commonly directs attention to the apparent scene rather than to intrinsic qualities of an inner display. Harman uses the observation to question a proposed argument against functionalism: that we directly apprehend qualities of experiences which could not be captured by descriptions of their functional role.

His point does not depend on the tree being real. A hallucination can represent a tree without either a real tree or a mental tree existing. Harman compares this with searching for the Fountain of Youth: failure of the supposed object to exist does not mean the explorer was actually searching for an idea. Intentional description does not turn a nonexistent worldly object into a second object inside the mind.

**Object Properties and Representation Properties**

A painting can depict a unicorn with four legs and a horn without having four legs or a horn itself. Likewise, an imagining or visual experience can represent a red object without being red in the ordinary sense.

Harman argues that several claims about qualia confuse properties attributed to an object with properties of the representation. The apparent redness of a tomato need not be an intrinsic red property of the experience. This leaves room for experiences to have psychological, neural, and representational properties; it denies the inference that every property presented through experience belongs to experience itself.

The distinction also prevents a simple caricature of representationalism. Harman is not saying that hallucinated objects must somehow exist outside the subject. He is explaining how a state can present things as being a certain way when no actual object fits that presentation.

**Functionalism and Three Objections**

Harman frames the paper as a defense of **psychophysical functionalism**, on which mental states are characterized through their causal relations to inputs, other states, and behavior. Its three connected targets are:

1. The claim that direct awareness discloses intrinsic experiential features which functionalism cannot accommodate.
2. The claim that complete knowledge of the functional role of vision still leaves a person ignorant of what seeing red is like.

3. The claim that spectrum inversion could preserve every functional relation while changing experience.

He argues that distinguishing represented properties from representation properties weakens the objections based on direct awareness and inversion. Inversion must be examined for changes in what and how a subject represents, rather than simply described as different inner color while everything else remains fixed.

His knowledge-argument reply challenges the completeness premise. A person blind from birth, he argues, lacks the full concept of redness that normally functions in visual perception. Without that concept, the person cannot know a particular functional fact: that seeing red normally produces a visual representation involving that very concept. Harman therefore denies that the person knows all the functional facts. He explicitly disagrees with David Lewis's ability-only reply, which grants complete information while claiming that experience adds abilities. Critics can contest Harman's account of concept possession and whether it genuinely defeats the completeness premise; it is not a claim that learning adds no information.

These are Harman's proposed responses, not settled conclusions of the debate. His paper does not establish a universal consensus that the knowledge argument or spectrum inversion has been refuted.

**Criticisms and Further Developments**

**Transparency versus identity.** Critics distinguish the phenomenological observation from a metaphysical conclusion. Even if introspection does not reveal intrinsic experiential properties, such properties might exist without being objects of attention. A representationalist needs a further argument for the claim that representational content exhausts phenomenal character. Harman's object/experience distinction undermines one purported source of evidence for intrinsic qualia but does not automatically decide every version of qualia realism.

**Blur and other visual phenomena.** Seeing a sharply bounded object blurrily seems different from seeing an object whose boundaries are themselves indistinct. Ned Block uses such cases to question whether all experiential differences are differences in represented scene. Representationalists, including Michael Tye, reply that blur can involve less determinate representation of boundaries rather than an extra internal property. Whether this captures the relevant contrast remains contested.

**Phenomenal knowledge.** Knowing what red is like appears to some critics to include a new truth about experience, not merely new abilities. Physicalist responses divide over this issue: ability hypotheses, new modes of presentation, and phenomenal-concept accounts should not be attributed to Harman as a single interchangeable solution.

G. E. Moore's earlier discussion of the apparent diaphanousness of sensation is a historical predecessor. Moore's own conclusions differ from Harman's, however, so transparency's history is not a continuous argument with an unchanged conclusion. Tye and Dretske subsequently developed more explicit representational theories with their own accounts of content and consciousness.

**Related Concepts and Figures**

- *Transparency* — the apparent accessibility of the scene rather than intrinsic mental qualities.
- *Intentionalism* — theories relating phenomenal character to representation.
- *Functionalism* — the theory Harman defends in the 1990 paper.
- *Mary's Room* — phenomenal knowledge and physicalist replies.
- *G. E. Moore* — an earlier discussion of the diaphanousness of sensation.
- *Michael Tye* — subsequent development of transparency arguments.
- *Ned Block* — challenges involving allegedly nonrepresentational experiential differences.

**Further Reading**


- Harman, Gilbert. "The Intrinsic Quality of Experience." *Philosophical Perspectives* 4 (1990): 31–52.
- Harman, Gilbert. *Reasoning, Meaning, and Mind*. Oxford University Press, 1999.


**Related Writing on This Site**

Mind, Matter, and Meaning presents the site author's arguments about mind and meaning. It is a separate argumentative work, not a neutral source for the account above.

**My View**

Harman gave the book one of its central moves. The transparency argument, as he formulated it via Eloise, reappears in my own words in the book's chapter on the transparency of experience: introspection delivers the world rather than an inner colored exhibit. I take the observation as powerful evidence for intentionalism and his object/experience diagnosis as fundamentally right. Transparency does not by itself prove strict identity; Chapter 6 supplies the further abductive case.

Where I do not simply follow Harman is on the larger program the transparency point was recruited to serve. His psychophysical functionalism is under-specified about what fixes content, what makes a content-bearing state conscious, and which persisting organization bears it. The book answers through externalism and producer-consumer history, a separate account of perceptual mode and the conscious threshold, and an integration-and-intervention test for the subject. Harman supplies the datum; the resulting synthesis belongs to the book.

I also do more defensive work than Harman's original paper does — answering Block's mental-paint cases (especially blurry vision, read as less determinate content) and conceding the pressure phosphene as genuinely hard. Holding Harman's conclusion while doing that additional work is the representationalist's actual position. See *Transparency*, *Michael Tye*, and *Ned Block*.

**Geoffrey Hinton**

Geoffrey Hinton is a computer scientist and cognitive psychologist whose research helped establish modern neural-network learning. His work on distributed representations, backpropagation, and deep learning also supports a distinctive account of thought and understanding. His philosophical claims appear partly in technical papers and partly in lectures and interviews; statements about machine experience must therefore be attributed to their particular context rather than treated as experimental findings established by his engineering work.

**Major Contributions and Sources**

- "Learning Distributed Representations of Concepts" (1986) explores learned representations of concepts and relations.
- "Learning representations by back-propagating errors," with David Rumelhart and Ronald Williams (1986), demonstrates learning of internal representations.
- The AlexNet work with Alex Krizhevsky and Ilya Sutskever (2012) contributed to advances in image recognition using deep convolutional networks.
- "The Forward-Forward Algorithm: Some Preliminary Investigations" (2022) proposes a different learning procedure and discusses hardware-dependent "mortal computation."
- "Artificial Intelligence: Turning our understanding of the mind right side up," a CIFAR lecture and discussion whose recording was uploaded in 2017, the 2025 Royal Institution lecture, and the 2026 Big Technology interview present arguments about meaning, subjective experience, and AI risk.

Hinton's role in backpropagation should not be described as its sole invention. In the June 2026 interview, he expressly notes that several groups developed the algorithm and emphasizes his collaborators' demonstration that it could learn interesting representations.

**Distributed Representations and Thought**

A distributed representation encodes information through a pattern of activity across many units rather than assigning each concept one isolated symbol or unit. Features can be shared across representations, allowing a system to learn similarities, differences, and relationships. Hinton treats this as a model of how brains acquire useful representations without receiving an explicit symbolic program.

In the CIFAR lecture recorded in the 2017 upload, he argues that vectors of neural activity constitute thoughts and that explicit symbol manipulation depends on a more basic representational system. This challenges classical pictures of thought as primarily the manipulation of sentence-like symbols. It does not imply that people never reason with explicit rules; it concerns the machinery and capacities on which such reasoning depends.

His examples combine learned semantic features with relations among words and concepts. A network can learn family relationships or infer the use of a new word by locating it within a structured system of features and relationships. Hinton interprets successful contextual prediction, analogy, and inference as evidence of understanding rather than as an alternative to understanding.

**Meaning and the System That Understands**

Hinton argues that a system understands language by constructing internal representations whose interactions support further cognitive work. On this view, the physical text does not itself contain understanding; the processing system must learn how its features fit together.

The discussion accompanying that lecture also characterizes mental states through normal causes and effects. A state associated with seeing red can be described through the red things that normally produce it. A desire can be described through the actions it tends to produce. These

causal descriptions supplement, but do not yet constitute a complete theory connecting, word-vector relationships with worldly reference.

His reply to Searle's Chinese Room is a **systems reply**. A person executing local instructions need not understand the resulting language, just as an individual neuron need not understand a sentence. Hinton argues that the capacities of an organized whole cannot be inferred from the absence of those capacities in a component. Searle contests the sufficiency of the system's formal processing; the exchange concerns the capacities and causal organization of the whole, not a requirement that every component understand.

**Subjective Experience and the Prism Example**

Hinton rejects an inner theater containing special mental objects. In his 2025 Royal Institution lecture, a report of hallucinated pink elephants describes how a perceptual system is representing the world incorrectly. The elephants characterize what would have to exist for that representation to be accurate; they need not be miniature pink objects in the head.

He then imagines a multimodal chatbot with a camera and an arm. It can point at an object. A prism shifts the apparent location, so the system points incorrectly. After learning about the prism, it reports that the object was really straight ahead although it had the subjective experience of seeing it elsewhere. Hinton argues that this use of experience language performs the same explanatory role it does for a person and therefore warrants attribution of subjective experience.

This is a thought-experiment-style argument, not a reported controlled demonstration that an actual robot was conscious. Hinton presents subjective experience as a starting point for a broader account of consciousness, which may involve additional capacities such as self-modeling.

In the June 2026 Big Technology interview, he explicitly says that he believes AI models are already conscious. He relates this to ordinary uses of awareness and to his rejection of the theater picture. That records his position in that interview; it does not establish a consensus about current AI consciousness.

**Implementation and Mortal Computation**

Hinton does not attribute consciousness to an abstract software description considered apart from implementation. His candidate is a system running in physical hardware. This permits artificial minds in principle without making implementation irrelevant.

"Mortal computation," discussed in the Forward-Forward paper, contrasts conventional digital computation with learning adapted to a particular, potentially irregular physical device. If learned parameters depend on that device's specific properties, knowledge may not transfer through simple copying of weights. Digital systems instead gain advantages from reproducibility and the ability of copies to share learned parameters. The proposal concerns learning, hardware, and energy use; it is not by itself a test for consciousness or personal identity.

**Objections and Replies**

**Understanding and reference.** Critics ask whether successful relations among representations establish what those representations refer to, especially for systems trained mainly on text.

Hinton's response stresses that extensive prediction requires learning the structures responsible for successful inference. The remaining dispute concerns whether this competence is sufficient, or must be supplemented by other causal, social, or perceptual relations.

**Report and phenomenal experience.** A system may detect perceptual error and report it flexibly. Critics distinguish these achievements from evidence that there is something it feels like to be that system. Hinton challenges the demand for a further private ingredient and asks what justifies withholding experience when the relevant functional explanation is shared. Rejecting an inner theater alone does not decide this disagreement: phenomenal realists need not believe in an internal spectator either.

**System boundaries and copies.** Attributing a mental capacity to a running system leaves open whether the relevant bearer is one execution, a persistent agent, or some larger organization. Copied weights do not by themselves settle whether there is one continuing subject or several successors. Hinton's hardware distinction makes these questions concrete without providing a settled account of identity.

His warnings about AI risk are separate from the consciousness claim. Highly capable systems might acquire harmful goals or instrumental strategies whether or not they have subjective experience. Consciousness, intelligence, controllability, and welfare should not be used as synonyms.

**Related Concepts and Figures**

- *Functionalism* — organizational accounts of mental capacities.
- *The Chinese Room* — the systems reply and its critics.
- *The Inner Theater* — Hinton's rejection of the inner-display model.
- *Information and Meaning* — representation, prediction, and understanding.
- *Simulation and Realization* — abstract models and physical implementations.
- *Daniel Dennett* — a related deflationary approach to consciousness.
- *John Searle* — an opposing account of understanding and computation.

**Further Reading**


- Hinton, Geoffrey E. "Learning Distributed Representations of Concepts." *Proceedings of the Eighth Annual Conference of the Cognitive Science Society* (1986): 1–12.
- Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." *Nature* 323 (1986): 533–536.
- Hinton, Geoffrey. "The Forward-Forward Algorithm: Some Preliminary Investigations." arXiv:2212.13345, 2022.
- Hinton, Geoffrey. "Artificial Intelligence: Turning our understanding of the mind right side up." CIFAR lecture and discussion. Recording uploaded by The Artificial Intelligence Channel, September 12, 2017. Video, especially 38:00–38:38 and 41:51–45:27.
- Hinton, Geoffrey. *Will AI outsmart human intelligence?* Royal Institution lecture, published July 22, 2025.
- Kantrowitz, Alex. Interview with Geoffrey Hinton. *Big Technology*, published June 4, 2026.

**Related Writing on This Site**

Mind, Matter, and Meaning presents the site author's arguments about mind and meaning. It is a separate argumentative work, not a neutral source for the account above.

**My View**

I think Hinton gets several important things right. Thought need not consist of sentences inscribed somewhere behind the eyes. Learned distributed representations can carry real causal structure, and analogical or intuitive competence need not reduce to explicit rule-following. His reply to Searle also lands: the ignorance of a component cannot decide the capacities of the organized whole. Most importantly, his attack on the inner theater runs beside this book's own. I applaud and agree completely with Hinton's view that perception does not place a private picture before an inner spectator, and hallucinated elephants do not acquire their pinkness from a pot of mental paint.

His account of experience nevertheless outruns its demonstration. The prism example shows a system registering a discrepancy between a perceptual representation and a corrected belief, then describing the representation in world-directed terms. That amounts to something substantial: perceptual content, error detection, metacognitive access, and flexible report. It does not yet establish that the relevant state has phenomenal character. Hinton moves from *using the words "subjective experience" as we do* to *having subjective experience as we do*. That inference works only if linguistic and functional equivalence already exhausts the phenomenon under dispute. His argument clears away a bad model of experience; it does not by itself identify what remains.

The same distinction tempers his account of meaning. Feature vectors explain how a system stores similarities, resolves ambiguity, supports analogy, and coordinates words within a learned model. They explain a great deal more than "mere statistics" suggests. Yet relations among internal vectors do not, by themselves, determine what those vectors refer to. Hinton's earlier appeal to normal worldly causes and effects points in the right direction without showing how that relation integrates with text-first learning. Training on human language can transmit public reference and install causally active derived content; multimodal perception and action can add world-involving anchors. Which bearer inherits which content depends on that history and organization, not simply on the geometry of an embedding space.

I therefore grant some present systems genuine representational competence and leave room for bounded understanding where inference, correction, learning, planning, and action constrain one another. I do not infer consciousness from fluent report, successful analogy, or the correct use of mental vocabulary. The running system rather than the abstract program supplies the right candidate bearer, as Hinton recognizes, but we still need evidence that a persisting artificial organization integrates perceptual content, recurrence, global availability, memory, self-modeling, and action into one cognitive economy. No carbon monopoly follows; neither does automatic admission for silicon.

Mortal computation sharpens questions about implementation, identity, energy use, and perhaps autonomy. It supplies no criterion of meaning or consciousness. Hinton's current verdict—AI is conscious—therefore exceeds the available evidence, although his challenge deserves better than

dismissal. Some artificial systems may already understand in limited ways. Whether any of them has a point of view remains open. And whether they threaten us remains a different question: an unconscious but highly capable agent could still ruin the afternoon.

### David Hume

David Hume (1711–1776) was a Scottish philosopher, historian, and essayist. His work examines how experience, habit, imagination, and feeling shape human thought and action. In philosophy of mind, he is associated with the distinction between impressions and ideas, the analysis of causal reasoning, and the account of the self as a bundle or succession of perceptions.

### Major Works

- *A Treatise of Human Nature* (1739–1740) investigates understanding, passions, and morals.
- *An Enquiry Concerning Human Understanding* (1748) reworks central themes concerning knowledge and causal inference.
- *An Enquiry Concerning the Principles of Morals* (1751) develops his moral philosophy.
- *The History of England* (1754–1762) established his reputation as a historian.
- *Dialogues Concerning Natural Religion* (1779, published posthumously) examines arguments about God through a philosophical dialogue.

### Impressions, Ideas, and Association

Hume divides perceptions into **impressions** and **ideas**. Impressions include sensations, passions, and emotions; ideas ordinarily have less force and vivacity and occur in memory and imagination. His copy principle says that simple ideas derive from corresponding simple impressions. Complex ideas can be recombined: imagining a golden mountain does not require previously seeing one.

This is not the claim that every thought is a literal internal photograph. Nor does it deny reasoning independent of particular observations. Hume distinguishes relations of ideas, such as mathematical relations, from matters of fact. His empiricist question concerns the materials and legitimate application of ideas, especially when apparently substantive metaphysical terms lack an identifiable experiential origin.

Hume himself raises the **missing shade of blue**: someone acquainted with neighboring shades might form the idea of a missing shade without having seen that precise shade. He treats it as an unusual exception rather than abandoning the copy principle. Its significance remains disputed, but a reference account should not present the principle as one for which he recognized no counterexample.

The principles of association—resemblance, contiguity, and causation—help explain how one perception brings another to mind. They provide a psychological account of connected thinking without requiring every transition to be a conscious inference.

### Causation, Induction, and Natural Belief

Observing one event followed by another does not reveal a necessary connection joining them. Repeated conjunction produces an expectation: fire comes to lead us to anticipate heat. Hume locates this experienced necessity partly in the mind's customary transition rather than in an independently perceived bond between events.

The problem of induction asks what justifies extending observed regularities to unobserved cases. A demonstration cannot establish that the future resembles the past, since a change is conceivable. Appealing to the previous success of induction already relies on the kind of inference needing justification.

Hume nevertheless does not propose that ordinary people stop making causal judgments. His explanation of habit and natural belief addresses why such reasoning persists. Interpreters disagree about whether his treatment makes causation merely a projection, permits realist commitments about unknown powers, or primarily limits what our ideas and evidence establish.

### The Bundle Account of the Self

In *Treatise* I.4.6, Hume reports that introspection encounters particular perceptions—heat, cold, pleasure, pain, love, hatred—but never an additional, unchanging self apart from them. The mind is therefore described as a bundle or succession of perceptions rather than as a directly apprehended substantial ego.

Resemblance, causal connection, and memory help explain why imagination attributes identity across changes. Memory reveals relations among perceptions and contributes to the belief in continuity; it does not make every remembered episode a separately sufficient test of identity.

Hume uses the theater metaphor but immediately qualifies it: there is no known theater or place containing the perceptions, and the comparison should not be taken to establish one. Reading the metaphor as his endorsement of an inner spectator reverses the role it plays in the argument.

In the Appendix to the *Treatise*, he acknowledges a difficulty in his account. Distinct perceptions and the lack of a perceived real connection among them leave the explanation of their unity unsatisfactory. Commentators disagree about the precise nature of this difficulty. It should not be reduced to a confident verdict that he secretly reinstated a soul or that one simple modern organizational theory completes his argument.

### Criticisms and Interpretive Questions

**The limits of introspection.** Failing to encounter an enduring self as one item among perceptions need not establish that there is no subject of those perceptions. Critics distinguish a subject from an object discoverable within experience. A Humean reply questions what content and evidence support a further substantial ego. The issue is whether such a subject requires the kind of impression Hume demands.

**Unity and continuity.** A collection of perceptions seems to need some account of why these perceptions belong together rather than to different minds. Hume's Appendix recognizes pressure at this point. Later bundle, psychological-continuity, and embodied theories propose different answers, but none should be simply attributed back to Hume.

**Skepticism and naturalism.** His doubts about rational foundations coexist with detailed explanations of the natural operations of thought. Emphasizing only skepticism misses the constructive psychology; emphasizing only naturalism risks treating psychological inevitability as a proof of truth. The relation between these strands remains central to interpretation.

His work on passions and morals further challenges the idea that reason alone determines motivation or ethical judgment. That wider project situates his theory of mind within an account of human life rather than an isolated theory of sensory images.

**Related Concepts and Figures**

- *The Self* — substantial, bundle, and continuity accounts.
- *Phenomenology* — descriptions of what introspection reveals.
- *The Inner Theater* — the limits of the theater metaphor.
- *John Locke* — empiricism and personal identity before Hume.
- *Immanuel Kant* — a response to Hume's questions about knowledge and causation.
- *Daniel Dennett* — a later account of selfhood without an inner owner.

**Further Reading**

- Hume, David. *A Treatise of Human Nature* (1739–1740), especially Book I and the Appendix.
- Hume, David. *An Enquiry Concerning Human Understanding* (1748).
- Hume, David. *An Enquiry Concerning the Principles of Morals* (1751).
- "David Hume." *Stanford Encyclopedia of Philosophy*.

**Related Writing on This Site**

Mind, Matter, and Meaning presents the site author's arguments about mind and meaning. It is a separate argumentative work, not a neutral source for the account above.

**My View**

Hume plays two roles in the book, and they pull in opposite directions. As an empiricist theorist of perception he *deepens* the inner theater Locke furnished: experience becomes a parade of impressions and their fainter copies, with the mind a kind of stage across which they pass. But as the philosopher who went looking for the self and *found no one home*, Hume hands the book one of its sharpest anti-reification weapons. I resist the first Hume and recruit the second.

Against the impressions-and-ideas framework: the copy theory of ideas reifies representational activity into inner pictures viewed by an inner eye. Memory is world-directed, not an album of fading impressions. Hume's picture is the empiricist inner theater at its most explicit.

From the bundle theory I take the anti-reification lesson: introspection turns up no additional inner owner behind experience. But a mere collection of perceptions does not yet identify their subject. The book's provisional answer is organizational: the minimal subject is the relatively maximal persisting control organization within which attention, memory, correction, expectation, planning, and action participate in one recurrent economy. Practical stake may add interests and welfare. Public language and inner speech arrive later, assembling the narrating personal self and its auditable ledger. Hume removes the tenant; Chapter 17 distinguishes the integrated subject from the socially assembled narrator.

**Frank Jackson**

Frank Jackson is an Australian philosopher whose work spans perception, metaphysics, language, and ethics. His knowledge argument became a prominent challenge to physicalism, and he later

rejected its antiphysicalist conclusion. An accurate account must distinguish the argument he originally defended from the representationalist physicalism developed in his later work.

**Major Works**

- *Perception: A Representative Theory* (1977) defends a sense-datum account of perception.
- "Epiphenomenal Qualia" (1982) introduces the knowledge argument and defends causally inert phenomenal properties.
- "What Mary Didn't Know" (1986) restates the argument and responds to objections.
- *From Metaphysics to Ethics* (1998) develops his method of conceptual analysis and a physicalist approach.
- "Mind and Illusion" (2003) explains his later rejection of the knowledge argument through representationalism.

**Mary and the Knowledge Argument**

Mary learns through black-and-white materials and has never experienced color. She nevertheless knows all physical facts about color vision. Jackson's stipulation includes not only neural events and wavelengths but also the relevant causal, relational, and functional facts. The case does not turn on the incompleteness of present neuroscience.

When Mary first sees red, Jackson originally argues, she learns what seeing red is like. If she already knew every physical fact, this new knowledge appears to reveal something nonphysical. The argument thus connects an epistemic premise about learning with a metaphysical conclusion about what facts exist.

In the 1986 paper, Jackson emphasizes that the missing knowledge concerns other people's experiences as well as Mary's own. Her release would show her that her prior understanding of their visual lives was incomplete. He also separates lack of knowledge from lack of imaginative ability: being unable to imagine an experience is not by itself the argument's conclusion.

**Early Epiphenomenalism and Perception**

In "Epiphenomenal Qualia," Jackson accepts nonphysical qualities while arguing that they need not affect bodily behavior. This **epiphenomenalism** avoids positing additional mental causes that interfere with physical causal explanations. It does not avoid every difficulty of dualism. Critics ask how causally inert qualities could explain our reports about them, our knowledge of them, or the reliability of those reports.

The earlier book *Perception* defended awareness of mind-dependent sense-data and a mediated relation to external objects. Arguments from illusion and perceptual variation supported that account. It should not be treated as interchangeable with his later representationalism: representing an object is not automatically perceiving a sense-datum that stands in for it.

**Later Physicalism and Representationalism**

"Mind and Illusion" rejects a response based only on the necessary a posteriori: Jackson does not regard saying that the experiential is physically necessitated but not a priori derivable as a satisfactory solution. He seeks an account under which the character of experience could, in principle, be deduced from a sufficiently rich physical account.

His later explanation distinguishes properties represented in experience from properties instantiated by the experience itself. Seeing red presents something as red; it need not acquaint the subject with an intrinsic nonphysical redness in the mind. Jackson argues that mistaking represented properties for instantiated experiential properties helps produce the antiphysicalist intuition.

He does not identify visual experience with any belief having the same simple content. A person can believe that a red, round object is present without seeing it. In "Mind and Illusion," perceptual representation has distinctive **richness, inextricability, immediacy, causal content, and functional role**. Color, shape, location, and movement are presented together rather than as detachable sentence parts. Sensory experience also affects the transition between beliefs without necessarily producing belief in the represented scene: an illusion can persist after a person knows it is an illusion.

Jackson argues that Mary acquires such a representational state on release. This gives her capacities to recognize, remember, and imagine that state. He agrees with Lewis and Nemirow's ability approach at this point, but reaches it through representationalism. He denies that she thereby learns a new proposition about a further property of the world. That is not the same as physicalist theories which grant new propositional knowledge under a new mode of presentation while denying any new nonphysical fact.

### Objections and Replies

**New knowledge or new abilities?** Critics regard Mary's first-person thought about seeing red as factual knowledge, not merely practical skill. Ability theorists distinguish possession of an experience-based ability from learning an additional fact. Other physicalists accept factual learning while disputing the inference to a nonphysical truthmaker. Jackson's later position is one member of this range, not the common position of every physicalist.

**The idealization.** Critics question whether anyone who knew every physical fact could really be ignorant in the stipulated way. Jackson's early response is that the stipulation isolates the explanatory issue rather than describing a feasible experiment. The disagreement concerns what complete physical knowledge includes, not whether a person can actually learn an infinite textbook.

**Representation and feel.** Critics of his mature theory ask whether the distinctive representational features suffice for phenomenal experience or merely describe its cognitive role. Jackson offers their combined operation as an explanation of feel. The success of that explanation depends on defending the representational account rather than invoking his change of mind as evidence for physicalism.

His reversal also does not refute the knowledge argument for other philosophers. Its premises and competing interpretations remain subjects of debate regardless of its originator's later judgment.

### Related Concepts and Figures

- *Mary's Room* — the knowledge argument and its competing interpretations.
- *Qualia* — the phenomenal properties at issue in early Jackson.
- *Physicalism* — the later position and alternative physicalist replies.

- *Intentionalism* — the mature representational theory.
- *Indirect Realism and Sense Data* — the earlier account of perception.
- *David Lewis* — the ability hypothesis.
- *Michael Tye* — other representational approaches to phenomenal knowledge.

**Further Reading**


- Jackson, Frank. "Epiphenomenal Qualia." *The Philosophical Quarterly* 32, no. 127 (1982): 127–136.
- Jackson, Frank. "What Mary Didn't Know." *The Journal of Philosophy* 83, no. 5 (1986): 291–295.
- Jackson, Frank. "Mind and Illusion." *Royal Institute of Philosophy Supplement* 53 (2003): 251–271.


**Related Writing on This Site**

Mind, Matter, and Meaning presents the site author's arguments about mind and meaning. It is a separate argumentative work, not a neutral source for the account above.

**My View**

Jackson is the figure whose *career* the book most admires, almost independently of any single argument. He built the cleanest anti-physicalist intuition pump of the late twentieth century, Mary's Room, and then, years later, decided it was wrong and came over to physicalism. The recantation is not a footnote; it is the book's preferred destination, reached by the very philosopher who built the best road away from it.

The book parts from *early* Jackson on two fronts. On perception, it rejects the representative theory of *Perception* (1977): sense-data are the inner intermediaries transparency and Austin remove. On consciousness, it rejects the move from Mary's epistemic novelty to a non-physical truthmaker. The ability hypothesis captures a real part of her change, but not all of it. The book grants that Mary gains a new true thought and reads it through an a-posteriori, acquaintance-and-recognition structure: one physical fact reached by descriptive and first-person routes.

Later Jackson is therefore an ally, not the book's exact destination. His recantation, physicalism, and strong representationalism all matter deeply, but his mature view denies the propositional novelty this book grants. The house Type-B account is more concessive about Mary's knowledge. It also keeps the phenomenal profile precise: externally fixed content under a perceptual mode, with determinacy, field structure, and attentional organization. World-involving history fixes the content; a separate state-consciousness relation supplies the domain. Chapter 8 states the negative result; Chapter 11 develops the acquaintance-recognition account.

**Immanuel Kant**

Immanuel Kant (1724–1804) was a German philosopher who worked in Königsberg. His critical philosophy investigates the conditions and limits of knowledge, moral agency, and judgment. In philosophy of mind, his accounts of sensibility, understanding, and self-consciousness offer an alternative both to knowledge founded entirely on sensory impressions and to metaphysics claiming knowledge beyond possible experience.

### Major Works

- *Critique of Pure Reason* (1781; substantially revised in 1787) examines knowledge, space and time, categories, and the limits of metaphysics.
- *Prolegomena to Any Future Metaphysics* (1783) provides another presentation of the critical project.
- *Critique of Practical Reason* (1788) addresses moral reason and freedom.
- *Critique of Judgment* (1790) examines aesthetic and teleological judgment.

### Intuitions and Concepts

Kant distinguishes **sensibility**, through which objects are given, from **understanding**, through which they are thought. An intuition presents a particular object; a concept supplies a general way of thinking about objects. Human knowledge requires their cooperation. Concepts without anything given do not establish knowledge of objects, while intuition alone does not supply the conceptual structure of an objectively valid judgment.

Receptivity and spontaneity therefore have different roles. Sensibility receives what affects the subject. Understanding combines representations according to rules. This is not a claim that people consciously invent every object they perceive. The question concerns the conditions under which representations can amount to experience of an objective world.

The **categories**, including substance and causality, are pure concepts of the understanding. Kant's Transcendental Deduction argues that they legitimately apply to objects of possible experience. Their validity does not authorize extending them without limit to whatever might exist independently of the conditions of our knowledge.

### Space, Time, and Transcendental Idealism

Kant argues that space and time are forms of human sensible intuition rather than features known to characterize things as they are in themselves. Objects encountered in space and time are **appearances**. We can have empirical knowledge of them without thereby knowing things considered independently of our manner of cognition.

**Transcendental idealism** is compatible, for Kant, with **empirical realism**: ordinary spatial objects need not be illusions or private mental pictures. His distinction concerns their status relative to the conditions of possible experience. It should not be presented simply as the view that the mind invents the external world.

Interpretation remains contested. Some readings understand appearances and things in themselves as distinct kinds of entities. Others treat the distinction as two aspects of the same objects: considered under our conditions of cognition and considered independently of those conditions. Epistemic readings emphasize limits on knowledge; more metaphysical readings ascribe stronger claims about what appears and what exists independently. These approaches disagree over both textual interpretation and philosophical viability.

"Thing in itself" and "noumenon" also require care. A noumenon in the negative sense concerns what is not an object of sensible intuition; a positive noumenon would require a kind of intel-

lectual intuition Kant denies to human knowers. Treating all these terms as the name of a hidden world behind an inner screen obscures the distinctions.

### Unity of Apperception and the Self

The possibility of attaching "I think" to one's representations expresses the **transcendental unity of apperception**. Representations must be capable of belonging together in one self-conscious cognitive activity if they are to constitute objective experience. This is a condition of their combination, not an additional item introspection discovers alongside sensations.

Kant's criticism of rational psychology denies that this formal unity proves an immaterial, simple, or immortal soul. The unity required for thought does not establish the substantial nature of the thinker. Likewise, empirical knowledge of oneself through inner sense remains subject to the conditions under which one appears to oneself.

This distinguishes his view from both a Cartesian inference to a thinking substance and a simple collection of unrelated impressions. The relation between conceptual synthesis, bodily life, and less reflective forms of awareness remains a subject of later development and criticism.

### Objections and Responses

**Affection and things in themselves.** If objects affect sensibility, critics ask how Kant can use a causal relation involving things in themselves while restricting the category of causality to experience. Different interpretations respond differently: some resist describing transcendental affection as an ordinary empirical causal claim, while others see a serious tension in the system. No single response commands agreement.

**Circularity and objective validity.** Critics ask whether the deduction of the categories presupposes the kind of unified objective experience it seeks to justify. Defenders distinguish an argument about necessary conditions from an attempt to derive experience without any starting point. Disputes concern which conditions the argument really establishes.

**Nonconceptual perception.** Later philosophers ask whether infants, animals, or fine-grained discrimination show that perception need not require the conceptual capacities Kant associates with objective cognition. Kant's requirements for judgment must be distinguished from claims about every sensory state. Contemporary conceptualists and nonconceptualists draw different lessons from this distinction.

**Idealism and realism.** Some critics reject the restriction of spatiotemporal knowledge to appearances. Kant's reply is that recognizing this restriction allows empirical knowledge while blocking unjustified metaphysical inferences. Whether that secures realism or limits it too severely depends partly on how transcendental idealism is understood.

### Related Concepts and Figures

- *The Myth of the Given* — later debates about sensibility and rational justification.
- *The Self* — unity of consciousness and knowledge of the subject.
- *Direct Realism* — disputes about appearances and perceptual access.
- *David Hume* — causation and the limits of empiricist explanation.

- *René Descartes* — rational psychology and certainty about the thinker.
- *John McDowell* — a later use of receptivity, spontaneity, and conceptual capacities.

**Further Reading**
- Kant, Immanuel. *Critique of Pure Reason* (1781/1787).
- Kant, Immanuel. *Prolegomena to Any Future Metaphysics* (1783).
- Kant, Immanuel. *Critique of Practical Reason* (1788).
- Kant, Immanuel. *Critique of Judgment* (1790).
- "Kant's Transcendental Idealism." *Stanford Encyclopedia of Philosophy*.

**Related Writing on This Site**
Mind, Matter, and Meaning presents the site author's arguments about mind and meaning. It is a separate argumentative work, not a neutral source for the account above.

**My View**
Kant functions here as a distant ancestor and a contested inheritance, not an ally. The current manuscript reaches its argument about singular reference chiefly through Strawson and Gareth Evans. It keeps Kant's contrast between concept and intuition and the thought that sensibility gives a finite thinker access to particulars; it leaves behind the claim that space and time come from the mind.

A second inheritance runs through Sellars and McDowell: experience cannot serve knowledge as a bare, self-certifying Given. I keep that lesson while declining McDowell's conceptualism. Perceptual content can retain a fine nonconceptual grain and still enter a rational, world-answerable economy. See *The Myth of the Given*.

The dividing line remains realism. Space and time belong to the world. Sensibility gives a creature a standpoint in that world; it does not author the frame or constitute its objects. The book therefore keeps a de-idealized Kantian insight about access while rejecting transcendental idealism. See *Direct Realism* and *Physicalism*.

**Ray Kurzweil**
Ray Kurzweil is an American inventor and writer whose books forecast developments in computing, artificial intelligence, biotechnology, and human enhancement. His account of mind emphasizes patterns of organization rather than the persistence of particular biological materials. The technological forecasts, claims about artificial consciousness, and arguments about personal identity are connected in his work but should be evaluated separately.

**Major Works**
- *The Age of Intelligent Machines* (1990) surveys machine intelligence and technological trends.
- *The Age of Spiritual Machines: When Computers Exceed Human Intelligence* (1999) explores future machine capacities and human responses to them.
- *The Singularity Is Near: When Humans Transcend Biology* (2005) develops the law of accelerating returns, human–machine integration, and patternism.

- *How to Create a Mind: The Secret of Human Thought Revealed* (2012) presents the Pattern Recognition Theory of Mind.
- *The Singularity Is Nearer: When We Merge with AI* (2024) revisits the forecasts and prospects for technological integration.

**Accelerating Returns and the Singularity**

Kurzweil's **law of accelerating returns** proposes that information technologies develop through compounding improvements and successive paradigms. A limit in one technology can be followed by a different technology that continues the broader trend. He uses historical curves to support forecasts of increasingly capable and inexpensive computation.

This is an empirical forecasting proposal, not a mathematical theorem guaranteeing indefinite exponential progress. Its application to intelligence, biotechnology, or economic transformation involves additional assumptions about what quantities can be measured and which constraints will become important.

Kurzweil has associated human-level machine intelligence with 2029 and the **Singularity** with approximately 2045. In his usage, the Singularity involves a transformation of human life through greatly expanded nonbiological intelligence, including integration with human intelligence. These are his forecasts, not agreed scientific dates or events already established by current systems.

**Patternism and Implementation**

Kurzweil calls his account **patternism**. An individual persists through an evolving pattern of organization even while bodily materials change. The account rejects identifying a person with a fixed inventory of atoms.

He argues that important features of a mind could in principle be implemented in another medium if its relevant organization were reproduced. This is stronger than the ordinary claim that computers can imitate some human performances. It connects reverse engineering of the brain with the prospect of recreating mental capacities.

The relevant pattern is not just an arbitrary string stored on a disk. The proposal concerns functioning organization instantiated in a medium. Determining how much structure must be reproduced—neural connectivity, dynamics, bodily interaction, or other features—is therefore central to whether an implementation could realize the capacities claimed.

**Uploading, Copying, and Personal Identity**

The 2005 book does not simply claim that every copy of a person's pattern is numerically that person. In its discussion "Who Am I? What Am I?", Kurzweil explicitly considers a copy that retains his memories while the original remains alive. He says that the copy would have its own experiences and would not straightforwardly be the original.

He contrasts this with gradual replacement of small parts of the brain by functionally equivalent components. This appears to preserve one uninterrupted life without producing a simultaneous original and copy. Yet he then asks why such replacement should differ fundamentally from

reconstruction, given the ordinary turnover of bodily material. He identifies a dilemma rather than supplying an uncomplicated proof that copying guarantees survival.

Kurzweil continues to affirm patternism despite these difficulties. The distinction between preserving psychological organization and preserving one person's numerical identity remains unresolved in the discussion. His hopes for uploading and radical life extension should therefore not be paraphrased as a settled identity theorem.

### Consciousness and Pattern Recognition

Kurzweil expects sufficiently developed artificial systems to display the subtle behavioral and expressive cues by which people ordinarily recognize consciousness. He argues that humans will come to accept such systems as conscious. At the same time, he acknowledges that access to these cues is not direct access to another system's subjective experience.

This keeps a distinction between public evidence and first-person reality within his account. Acceptance by others, behavioral equivalence, and consciousness are not identical claims, even if he expects them to converge in practice.

The **Pattern Recognition Theory of Mind** presents the neocortex as a hierarchy of pattern recognizers. Lower-level patterns combine into higher-level ones, supporting recognition, prediction, and other cognitive capacities. The theory motivates engineering proposals as well as an account of biological cognition. Whether that architecture adequately captures the diversity of brain mechanisms, and whether implementing it would establish consciousness, are further questions.

### Criticisms and Responses

**Extrapolation.** Critics ask whether selected technological curves justify predictions about general intelligence or broad social change. Energy requirements, physical limits, data, institutions, and uneven progress may interrupt an extrapolation. Kurzweil's response emphasizes transitions between technological paradigms rather than indefinite improvement in one device type. That response still needs evidence that the proposed transitions are feasible and relevant to each forecast.

**Simulation and realization.** A model of neural activity may reproduce a formal pattern without reproducing every causally relevant feature. Computational functionalists argue that the right organization could suffice across substrates; other accounts require additional implementation-specific properties. The Church–Turing thesis does not by itself decide this question, since a thesis about computable functions is not a theory of consciousness.

**Continuation and duplication.** Two copies can share memories at creation and then diverge. This challenges an account equating shared pattern with one identity. Kurzweil recognizes the challenge himself; defenders of gradual replacement emphasize causal continuity, while other views separate survival of psychological traits from numerical identity.

These objections do not establish that artificial minds are impossible, and progress in artificial intelligence does not automatically resolve them. Forecast accuracy, mental capacity, consciousness, and personal continuation require their own evidence.

**Related Concepts and Figures**

- *Functionalism* — mentality and functional organization.
- *Simulation and Realization* — reproducing a model versus realizing a capacity.
- *Computation and the Church-Turing Thesis* — computability and its limits as a theory of mind.
- *The Turing Test* — behavioral evidence for machine intelligence.
- *The Self* — continuity, copies, and personal identity.
- *David Chalmers* — arguments about organizational invariance and consciousness.

**Further Reading**

- Kurzweil, Ray. *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*. 1999.
- Kurzweil, Ray. *The Singularity Is Near: When Humans Transcend Biology*. 2005.
- Kurzweil, Ray. *How to Create a Mind: The Secret of Human Thought Revealed*. Viking, 2012.
- Kurzweil, Ray. *The Singularity Is Nearer: When We Merge with AI*. Viking, 2024.

**Related Writing on This Site**

Mind, Matter, and Meaning presents the site author's arguments about mind and meaning. It is a separate argumentative work, not a neutral source for the account above.

**My View**

Kurzweil earns his place on this page by being right about what most people argue over, and wrong, I think, about something almost no one stops to examine. He is very likely right that machine competence will keep climbing, and right that nothing in the matter of the brain is sacred — the atoms do turn over, and a view of mind that pinned identity to particular carbon would deserve the patternist's scorn. I have no wish to defend substance over pattern. My disagreement is about what kind of pattern a self is.

Patternism treats a mind as a *type* — an organization of information, repeatable in principle, the same pattern whether instanced once or a thousand times. The book separates that proposal into several questions. Content depends on history and use; understanding depends on integrated content-guided capacities; consciousness requires the further phenomenal organization argued for in Chapter 6. Personal continuation poses another problem. Copying a complete pattern may create a successor that remembers the original life without making two later tokens numerically identical to one earlier person. Practical stake can make the loss matter to a token, but it neither creates the mind nor decides identity by itself.

The Pattern Recognition Theory of Mind makes a related compression one floor down. A hierarchy of pattern recognizers can support remarkable competence. Whether its states have original or derived content depends on their history; whether several capacities compose understanding depends on the bearer and its integrated use of those contents. Autonomy and welfare enter separately. Kurzweil offers a powerful architecture of competence and treats it too quickly as the whole of a mind.

None of this is the carbon chauvinist's complaint that only meat can think. A computational artificial system remains eligible in principle if its concrete history and organization establish the capacities claimed. A restore may count as a successor rather than the original continued, but that

verdict concerns token identity, not whether the restored system can mean, understand, feel, or acquire interests of its own. Kurzweil wants portability to settle every question at once. The book insists on asking them separately. See *Simulation and Realization*, *Functionalism*, and *Alan Turing*.

### Joseph Levine

Joseph Levine is an American philosopher whose work on consciousness focuses on the **explanatory gap** between physical explanations and phenomenal character. His 1983 paper gives a precise form to the concern that identifying an experience with a physical state may leave its felt character unexplained. The argument raises a problem for materialism without claiming to demonstrate that every psychophysical identity is false.

### Major Works

- "Materialism and Qualia: The Explanatory Gap" (1983) introduces the gap by reconsidering Kripke's argument against psychophysical identity.
- "Qualia: Intrinsic, Relational, or What?" (1995) examines competing characterizations of phenomenal properties.
- *Purple Haze: The Puzzle of Consciousness* (2001) develops the difficulty at book length, including questions about phenomenal concepts and first-person thought.

### The Explanatory Gap

Levine compares two proposed identities: heat with molecular motion, and pain with C-fiber firing. These are philosophical examples; the latter is not a claim that current neuroscience identifies every pain with one fiber type.

For heat, the physical explanation makes intelligible how molecular motion performs the causal roles associated with heating, expansion, and related effects. For pain, a physical and functional account may explain responses to injury, avoidance, and verbal reports. Levine asks why this account does not similarly make the felt quality of pain intelligible. Why should this state feel like pain rather than like something else, or have any felt character?

The question does not demand that a verbal explanation cause the reader to undergo the experience. It asks whether the connection between physical nature and phenomenal character can be explained rather than left as a brute pairing. A complete account of behavior may leave this demand unanswered.

Color supplies another example. Descriptions of the mechanisms associated with seeing red and green may explain discrimination and response. Levine argues that the descriptions still do not explain why the respective mechanisms yield those particular qualitative experiences. Apparently conceivable rearrangements help bring the explanatory deficit into view.

### Epistemology and Metaphysics

Kripke's challenge uses the apparent contingency of psychophysical identities to argue that they cannot be true necessary identities. Levine redirects the argument: the persistent intuition supports an epistemological problem rather than, by itself, a proof of nonidentity.

A physical identity might be true even when a thinker cannot make its truth intelligible from the physical description. The gap therefore concerns explanation and knowledge. This does not license the stronger conclusion that Levine has shown the gap to be merely superficial or fully explained by our conceptual limitations. In the 1983 paper, the explanatory problem remains serious enough to impede determining which psychophysical identities are true.

Likewise, saying that a zombie or inverted experience is conceivable does not alone establish its metaphysical possibility. Physicalists and their critics disagree about what further premises connect those notions. Levine's distinction leaves room for physicalism without treating the gap as harmless.

### Phenomenal Concepts and Later Work

Phenomenal concepts are ways of thinking about experiences through their phenomenal character—for example, thinking of a familiar kind of pain as that kind of feeling. They appear to differ from third-person physical descriptions even when both purport to concern the same state.

In the debate developed around *Purple Haze*, such concepts are candidates for explaining why physical information does not transparently reveal phenomenal truths. But a proposal that they refer to physical states is not yet an explanation of their distinctive first-person operation. Levine treats the puzzle of how experience enters thought as a further philosophical problem, not simply as a solved bridge from two concepts to one fact.

This distinguishes his concern from both an immediate inference to dualism and an uncomplicated phenomenal-concept strategy. A theory might account for why the concepts are different while still leaving unexplained why the referents have phenomenal character. Whether a conceptual explanation can meet the original demand remains disputed.

### Responses and Objections

**Functional and representational explanations.** Some theories identify phenomenal character with a sufficiently specified functional or representational organization. They argue that nothing further needs explaining once the phenomenon is correctly characterized. Levine's challenge is to show why that characterization captures felt character rather than only its roles or contents. The success of the reply depends on an independent defense of the proposed identity.

**A posteriori identities.** Another response accepts identities discovered empirically without requiring an a priori derivation between the relevant concepts. Levine grants that epistemic distance does not establish metaphysical distance. His objection is that consciousness still differs explanatorily from standard examples such as heat: the identity may remain intelligible only as an unexplained fact.

**Phenomenal-concept strategies.** These explain apparent separability through distinct modes of thinking about one physical state. Critics ask whether explaining the appearance of a gap also explains the phenomenon across the gap. Levine's work presses this distinction; calling the gap "conceptual" cannot function as a substitute for the needed account.

**Antiphysicalist uses of the gap.** Dualists may argue that the failure of physical explanation reveals something missing from the physical world. Levine's original argument does not entail

that conclusion. Additional premises about the relationship between intelligibility, conceivability, and possibility are required.

The explanatory gap is thus a shared problem around which theories differ. Its persistence does not settle whether physicalism is true, and a commitment to physicalism does not by itself remove it.

**Related Concepts and Figures**

- *The Explanatory Gap* — the distinction between identity and intelligible explanation.
- *The Hard Problem of Consciousness* — related questions about why physical processes involve experience.
- *The Phenomenal Concept Strategy* — explanations based on distinctive first-person concepts.
- *Physicalism* — metaphysical identity and explanatory adequacy.
- *Frank Jackson* — phenomenal knowledge and the knowledge argument.
- *David Chalmers* — further arguments from epistemic to metaphysical gaps.

**Further Reading**


- Levine, Joseph. "Materialism and Qualia: The Explanatory Gap." *Pacific Philosophical Quarterly* 64 (1983): 354–361.
- Levine, Joseph. *Purple Haze: The Puzzle of Consciousness*. Oxford University Press, 2001.


**Related Writing on This Site**

Mind, Matter, and Meaning presents the site author's arguments about mind and meaning. It is a separate argumentative work, not a neutral source for the account above.

**My View**

Levine names the problem the book cares most about and stands as a fellow-traveler on its diagnosis. He turns the anti-materialist pressure into an epistemological claim: even if a psychophysical identity holds, the physical description does not make the felt character intelligible. The gap remains real without yet becoming a second feature of reality.

The book adds two claims Levine's diagnosis does not supply by itself. First, the acquaintance-recognition account explains why description cannot confer the first-person recognitional capacity acquired through undergoing an experience. The two routes can remain conceptually isolated while reaching one physical fact. Second, the identity claim holds that phenomenal character *consists in* the complete world-presenting intentional profile. That claim closes the ontological gap only if the independent argument for it succeeds; acquaintance explains the epistemic residue but cannot manufacture the identity.

Levine therefore supplies the right problem and the right restraint. Chalmers presses the further inference from an epistemic gap to an ontological one; the book declines that inference. See *The Hard Problem of Consciousness* and *Physicalism*.

**David Lewis**

David Lewis (1941–2001) was an American philosopher whose work ranged across metaphysics, language, logic, and philosophy of mind. His accounts of mental states combine functional

characterization with identification of the states realizing those roles. His ability hypothesis offers a physicalist response to the claim that experience teaches facts unavailable through complete physical knowledge.

### Major Works in Philosophy of Mind

- "An Argument for the Identity Theory" (1966) argues from the causal roles of mental states to psychophysical identification.
- "Psychophysical and Theoretical Identifications" (1972) develops the use of implicit theoretical definitions.
- "Mad Pain and Martian Pain" (1980) examines atypical human pain and differently realized extraterrestrial pain.
- "What Experience Teaches" (1988) defends the ability hypothesis concerning phenomenal knowledge.

His wider metaphysical work includes *On the Plurality of Worlds* (1986). The arguments about mind should not be treated as dependent on accepting his entire theory of concrete possible worlds.

### Functional Roles and Their Realizers

Lewis begins with a network of commonsense psychological claims connecting beliefs, desires, sensations, stimuli, and behavior. Mental terms are treated as implicitly defined by their roles within this theory. Replacing the terms with variables allows the theory to characterize whatever states jointly occupy those roles.

A distinction is essential: a **role** is a causal specification; a **realizer** is the state that fills it. Lewis's identity approach identifies mental states with the relevant realizers, not simply with abstract roles. Empirical inquiry can then discover which physical states perform the roles in an appropriate population. The functional characterization and physical identification do different work.

This allows a mind–brain identity theory without requiring one physiological type for every possible creature. What normally realizes pain in humans may differ from what realizes it in another species. Multiple realization does not, on Lewis's account, erase the importance of the realizing state.

### Mad Pain and Martian Pain

The mad-pain case concerns someone whose pain has unusual causes and effects. It need not lead to the ordinary avoidance behavior or result from typical injury. If pain were determined solely by the individual's present functional role, the case would threaten the claim that this person is in pain.

The Martian case supplies the opposite pressure. A creature's pain has familiar causes and effects but is realized by hydraulic mechanisms rather than human neural tissue. An unqualified species-independent identity with one human neural state would exclude the Martian.

Lewis relativizes realization to appropriate populations. The atypical human can have the state that normally plays the pain role in humans, even though it plays an unusual role in that indi-

vidual. The Martian can have a different physiological state that normally plays the corresponding role in Martians. Atypicality within a population and variation between populations therefore receive different treatment.

Critics ask how to identify the relevant population, especially for unique artificial systems or borderline cases. Lewis's proposal provides a strategy for accommodating variation, not a mechanical procedure that removes every question about pain attribution.

### What Experience Teaches

The knowledge argument imagines Mary possessing every physical fact about color vision while having no color experience. On release, she seems to learn what seeing red is like. The antiphysicalist conclusion treats this as discovery of an additional nonphysical fact.

Lewis, following Laurence Nemirow, proposes that knowing what an experience is like consists in abilities to remember, imagine, and recognize it. Tasting Vegemite for the first time can enable a person to recall its taste, imagine it later, and recognize a recurrence. A description need not confer these abilities merely by stating all the relevant facts.

Lewis distinguishes **phenomenal information** from these acquired capacities. He does not deny that experience can also teach ordinary facts: someone may learn that a particular food tastes a certain way or recognize which food they were given. His claim is that the specifically phenomenal gain need not consist in information concerning a further property beyond the physical facts.

The ability hypothesis therefore accepts a substantial change in Mary while rejecting the interpretation required for the knowledge argument. It does not claim that experience teaches nothing or that an ability is itself a mysterious nonphysical addition.

### Objections and Replies

**Propositional learning.** Critics argue that Mary gains a true thought about how an experience feels, not just new skills. Ability theorists ask whether this thought expresses a new fact or deploys the newly acquired capacities. Other physicalists accept new propositional knowledge while holding that it concerns an old physical fact under a new mode of presentation. That alternative should not be equated with Lewis's hypothesis.

**Ability dissociations.** Imagination, recognition, and memory can vary independently. A person may know an experience while having limited ability to imagine it. This challenges an overly strict list of jointly necessary abilities. Defenses refine the relevant ability or distinguish normal manifestations from the underlying competence. Such refinements require argument; calling all phenomenal knowledge "know-how" does not alone establish them.

**Inverted and absent experience.** Critics of functional theories argue that a system might occupy the roles while feeling differently, or feel nothing. Lewis's account addresses some atypical and cross-species cases through population-relative realizers. Whether it captures every phenomenal contrast remains disputed.

**Folk psychology.** The initial theory might contain errors or fail to uniquely identify a set of occupants. This threatens implicit definition unless the account allows sufficiently accurate

approximations or revisions. The strength of the approach is its systematic treatment of interrelated mental terms; its vulnerability is dependence on the adequacy of that system.

**Related Concepts and Figures**

- *Functionalism* — roles and physical realizers.
- *Identity Theory* — psychophysical identification.
- *Mary's Room* — the knowledge argument and the ability hypothesis.
- *Qualia* — phenomenal character and functional characterization.
- *Frank Jackson* — the originator of the Mary argument.
- *Ned Block* — challenges to functional accounts of consciousness.

**Further Reading**


- Lewis, David. "An Argument for the Identity Theory." *The Journal of Philosophy* 63 (1966): 17–25.
- Lewis, David. "Psychophysical and Theoretical Identifications." *Australasian Journal of Philosophy* 50 (1972): 249–258.
- Lewis, David. "Mad Pain and Martian Pain." In *Readings in Philosophy of Psychology*, vol. 1, edited by Ned Block. Harvard University Press, 1980.
- Lewis, David. "What Experience Teaches." *Proceedings of the Russellian Society* (1988); reprinted in *Papers in Metaphysics and Epistemology*. Cambridge University Press, 1999, pp. 262–290.


**Related Writing on This Site**

Mind, Matter, and Meaning presents the site author's arguments about mind and meaning. It is a separate argumentative work, not a neutral source for the account above.

**My View**

Lewis enters the book at two important joints: Mary's Room and the functionalist tradition. His ability hypothesis identifies a real component of Mary's change. She gains powers to recognize, imagine, and remember red. Those abilities show that experience can teach without adding a new item to the world's inventory, but the current manuscript does not treat know-how as the whole of her gain.

The house reply grants Mary a genuinely new true thought and asks what makes it true. Acquaintance with the occurring state supplies a first-person demonstrative anchor and grounds a recognitional capacity that description alone cannot confer. The ability account describes part of that new equipment; the one-fact/two-routes account explains how epistemic novelty can remain ontologically conservative. The identity claim then supplies the metaphysical thesis on which the deflation depends.

Lewis's functionalism remains close but incomplete. Internally specified causal role does not determine content; world-involving history and producer-consumer organization help do that work. Wide organization may preserve the functionalist insight once the world returns to the account, but Lewis's narrow machinery cannot settle content by itself. See *Mary's Room*, *Functionalism*, and *Physicalism*.

### John Locke

John Locke (1632–1704) was an English philosopher whose work in epistemology, politics, education, and religion influenced later empiricism and liberal political thought. *An Essay Concerning Human Understanding* investigates the origins, limits, and degrees of human knowledge. Its accounts of ideas, perception, and personal identity remain central to debates about how a mind relates to the world.

### Major Works

- *An Essay Concerning Human Understanding* (1689) examines ideas, knowledge, language, and judgment; its chapter on personal identity was added in the second edition of 1694.
- *Two Treatises of Government* (1689) develops an account of political authority, rights, and legitimate resistance.
- *A Letter Concerning Toleration* (1689) argues for religious toleration within limits that also remain subjects of criticism.
- *Some Thoughts Concerning Education* (1693) addresses education and character formation.

### Sensation, Reflection, and the Way of Ideas

Locke rejects innate ideas and principles as an explanation of the contents of thought. Ideas arise through **sensation**, concerning external things, and **reflection**, concerning the mind's own operations, such as thinking, willing, and doubting. Simple ideas are received; the understanding can combine, compare, and abstract from them to form more complex ideas. This makes the mind more than a passive receptacle even within his empiricist account.

His definition of an idea as whatever is the object of the understanding when a person thinks leaves interpretive questions. A traditional **representative** reading makes ideas the immediate objects of awareness, through which external objects are known. On a strong "veil of perception" reading, the perceiver inspects inner items while the external world remains behind them.

Other interpreters resist this picture, treating ideas in terms of acts or contents of perception rather than a separate inner object. Scholars also distinguish accepting representation from accepting the veil's skeptical consequences. Locke's criticism of Malebranche and his confidence in sensitive knowledge of external things complicate attributing the strongest veil account to him. The dispute concerns what an idea is and what follows from its role, not whether Locke used the vocabulary of ideas.

### Primary and Secondary Qualities

In *Essay* II.8, **primary qualities** include solidity, extension, figure, motion or rest, and number. Locke regards these as qualities belonging to bodies, and he says our ideas resemble them. **Secondary qualities**, such as color, sound, and taste, are powers of bodies to produce ideas in perceivers through their primary-quality constitution.

A rose's color, on this account, depends on the body's power to affect an appropriate perceiver. The idea of redness does not resemble the particle structure responsible for that effect. Secondary qualities are therefore not simply nonexistent: bodies have the relevant powers. The account distinguishes those powers from the sensory ideas they produce.

Critics question the resemblance claim, the classification of particular qualities, and the grounds for knowledge of unobserved particles. Some also question whether primary qualities are less dependent on perceptual conditions than secondary ones. Locke's corpuscular explanation supplies a scientific framework for the distinction, while his repeated acknowledgments of limited knowledge constrain how much that framework can establish.

### Memory, Knowledge, and Mental Imagery

Locke compares the understanding with a dark room receiving pictures and memory with a storehouse. These metaphors have encouraged later inner-picture readings. They do not by themselves decide whether he believed in a literal internal spectator or collection of preserved mental objects.

His account of memory in *Essay* II.10 explicitly distinguishes retaining an idea from having it continuously present. Remembering involves reviving a perception with awareness of having had it before. Describing memory as a storehouse must therefore be read alongside his account of what retention actually involves, not treated as a complete mechanical theory.

Locke defines knowledge through perception of agreement or disagreement among ideas and distinguishes intuitive, demonstrative, and sensitive knowledge. Sensitive knowledge of presently existing external things has a different degree and basis of certainty from intuition and demonstration, but Locke does not simply dismiss it as opinion. Much ordinary judgment proceeds on probability rather than knowledge in this strict sense.

### Persons, Consciousness, and Identity

The chapter "Of Identity and Diversity" distinguishes the persistence of masses of matter, living organisms, human beings, and persons. A living body can remain the same organism while its material parts change. Personal identity requires a different inquiry.

A **person** is a thinking, intelligent being capable of reason, reflection, and considering itself as the same across times and places. Locke locates personal identity in sameness of consciousness, rather than in the sameness of a particular body or soul. In the prince-and-cobbler example, the prince's consciousness accompanies his soul into the cobbler's body. Locke's verdict distinguishes the same person from the same human being.

Memory extends consciousness backward, but interpreting consciousness as nothing more than direct episodic memory is contested. Some readings emphasize a continuing conscious life; others emphasize appropriation of past actions as one's own. Locke also calls "person" a **forensic term**, connecting selfhood with accountability, reward, and punishment. These concerns belong to the theory rather than being merely later applications of it.

### Objections and Interpretive Responses

**Circularity.** Joseph Butler argues that consciousness of identity presupposes identity and therefore cannot constitute it. Later responses distinguish a criterion of identity from evidence for it, or seek a psychological relation describable without already assuming one person. Whether these reconstruct Locke or state his original account remains disputed.

**Transitivity and forgetting.** Thomas Reid's brave-officer case imagines an officer remembering a childhood punishment and an old general remembering his life as the officer but not the childhood event. A direct-memory criterion threatens the transitivity of identity. Later psychological-continuity theories use overlapping chains rather than requiring direct memory of every earlier event. Interpreters of Locke also ask whether the strict memory reading is adequate in the first place.

**The veil problem.** If only inner ideas are known directly, critics ask how resemblance to an external original can be checked. Representative readings must explain how causal relations support external knowledge without comparing two independently inspected items. Alternative readings deny that Locke's theory has the obstructing inner object assumed by the criticism.

Locke's historical importance does not settle these disputes. His empiricism, limited account of substance, and distinction between persons and organisms continue to motivate different philosophical reconstructions.

### Related Concepts and Figures

- *Indirect Realism and Sense Data* — representative readings and their alternatives.
- *The Inner Theater* — later interpretations of mental-picture metaphors.
- *Direct Realism* — disputes about immediate perceptual access.
- *The Self* — consciousness and the persistence of persons.
- *René Descartes* — rationalist and substance-centered accounts.
- *David Hume* — later empiricism and criticism of substantial selfhood.
- *John McDowell* — a later debate about experience and knowledge.

### Further Reading

- Locke, John. *An Essay Concerning Human Understanding* (first published 1689), especially II.8, II.10, II.27, and Book IV. Use an edition containing II.27, "Of Identity and Diversity," added in the second edition of 1694.
- Locke, John. *Two Treatises of Government* (1689).
- "John Locke." *Stanford Encyclopedia of Philosophy*.
- "Locke on Personal Identity." *Stanford Encyclopedia of Philosophy*.

### Related Writing on This Site

Mind, Matter, and Meaning presents the site author's arguments about mind and meaning. It is a separate argumentative work, not a neutral source for the account above.

### My View

Locke matters to the book chiefly through the interpretation his *Essay* set loose. The influential Cartesian-Lockean reading furnished the inner theater with ideas, relocated secondary qualities toward the perceiver, and helped later philosophers and scientists treat inner representations as the immediate objects of awareness. That reception, rather than an uncontested verdict on Locke himself, forms the book's target.

The structural mistake runs from *we perceive by means of representational states* to *those states are what we perceive*. Causal mediation does not make the vehicle of perception its object. Austin's

attack on the sense-datum descendants of the view and the transparency of experience supply two routes back to the world. The historical caveat matters: if Locke's ideas name acts or contents of perceiving, the sharpest form of the charge lands on the veil tradition built from him rather than on Locke alone.

The book keeps Locke's naturalism and the empiricist insight that traffic with the world shapes a mind. Semantic externalism carries that insight farther by locating content in world-involving history while discarding the inner intermediary. See *Indirect Realism and Sense Data*, *The Inner Theater*, *Direct Realism*, and *Qualia*.

### John McDowell

John McDowell is a philosopher whose work spans perception, knowledge, language, ethics, and the interpretation of Aristotle, Kant, and Wittgenstein. *Mind and World* develops an account of how experience can rationally constrain thought without functioning as an uninterpreted foundation. His later writings revise aspects of that account while retaining its central concern with the world's availability to a rational subject.

### Major Works

- "Singular Thought and the Extent of Inner Space" (1986) examines thought, externalism, and the Cartesian conception of an inner standpoint.
- *Mind and World* (1994; paperback with a new introduction, 1996) develops the relation between receptivity and spontaneity, conceptualism, and second nature.
- *Mind, Value, and Reality* (1998) collects essays on ethics, perception, and naturalism.
- *Meaning, Knowledge, and Reality* (1998) collects work on language and epistemology.
- *Having the World in View: Essays on Kant, Hegel, and Sellars* (2009) includes later discussions of perceptual content, including "Avoiding the Myth of the Given."

### Experience Between the Given and Coherentism

McDowell wants empirical thought to answer to the world. Experience must be more than an external cause of belief: it must be able to supply reasons. Yet an appeal to a wholly nonconceptual Given seems to demand that something outside the space of reasons nevertheless justify what occurs within it.

He describes an oscillation. Foundationalism seeks rational constraint from bare sensory impacts. A coherentist response rejects that appeal but risks leaving beliefs answerable only to other beliefs. McDowell proposes that experience already involves conceptual capacities, so it can rationally constrain judgment without being a bare input to which thought later adds meaning.

This does not mean that the subject must endorse every appearance. A familiar visual illusion can persist after the subject recognizes it as misleading. Experience can present things a certain way while judgment declines to accept them. Conceptual involvement in perception must therefore be distinguished from freely forming a belief.

### Receptivity, Spontaneity, and Conceptual Content

Taking up Kant's terminology, McDowell distinguishes **receptivity**, in which things are given to the subject, from **spontaneity**, the capacity exercised in thinking and judging. His account

allows conceptual capacities to be passively actualized in experience. The subject does not have to choose how things appear for the appearance to belong to the space of reasons.

In *Mind and World*, experience is described as having content of a kind that can also be the content of a judgment: that things are thus and so. In a veridical case, this content is also how the world is. McDowell rejects the picture in which thought confronts an inherently unintelligible world through a separate inner intermediary.

His later account does not retain every part of this formulation. In "Avoiding the Myth of the Given," he distinguishes **intuitional** from **propositional** content. Experience can possess a unified, nondiscursive content from which judgments draw, without itself having the articulated form of those judgments. He continues to regard experience as an actualization of conceptual capacities. The revision should not be described either as no change at all or as a straightforward conversion to nonconceptualism.

### Second Nature and Naturalism

**Second nature** describes capacities acquired through upbringing, education, and participation in a linguistic community. Responsiveness to reasons becomes part of the life of a human animal rather than a supernatural power added from outside nature.

McDowell rejects an equation of everything natural with everything describable solely through the laws studied by natural science. This lets him acknowledge the legitimacy of scientific explanation while resisting reduction of rational relations to those explanations. His contrast between "bald naturalism" and "naturalized platonism" concerns how to understand nature and reasons, not a denial that humans are biological beings.

This perspective also informs his ethics. Moral sensitivity can be cultivated as part of a person's practical understanding. It need not be modeled exclusively as applying a detached rule to an independently given inventory of value-free facts.

### Disjunctivism and Openness to the World

McDowell's disjunctive treatment of perception rejects deriving the nature of a successful perception from what it shares with a subjectively indistinguishable hallucination. In the successful case, a fact about the world is perceptually available. In the deceptive case, it merely appears to be available.

From the subject's inability always to discriminate the cases, it does not follow that both provide only the same world-neutral evidence. This matters especially for skepticism: a genuine perception can supply an epistemic standing not possessed by its deceptive counterpart.

McDowell's epistemological motivations should be distinguished from every metaphysical version of naive realism. Disjunctivists differ over objects, facts, phenomenal character, and the explanation of hallucination. His position should not be presented as automatically interchangeable with M. G. F. Martin's theory or as a ban on scientific explanation of perceptual mechanisms.

### Philosophical Method and Criticisms

**Quietism and explanation.** McDowell often seeks to remove assumptions that make constructive philosophical theories appear compulsory. Critics ask whether this leaves the origins of content and rationality insufficiently explained. His reply distinguishes a legitimate scientific investigation from a philosophical demand generated by an overly restrictive conception of nature. Not every request for empirical explanation is his target.

**Nonconceptual perception.** Critics cite infants, animals, and the fine detail of perception as challenges to conceptualism. McDowell distinguishes rational human experience from other forms of animal sensitivity; he does not deny that animals perceive. His earlier response to fine grain invokes demonstrative capacities, such as thinking of "that shade," while later intuitional content allows more than an articulated judgment captures. Critics question whether these responses still require concepts where sensory discrimination does not.

**Rational constraint.** If conceptual capacities already operate in experience, critics ask whether the account really supplies a constraint independent of thought. McDowell emphasizes passivity and the factive character of successful perception: the world determines what is available, rather than the subject choosing the content. Whether this avoids idealism and provides an adequate response to skepticism remains contested.

**Hallucination and evidence.** A common-factor theorist asks what explains the cases' subjective indistinguishability. McDowell denies that this explanatory task requires giving the deceptive case the same epistemic status as the successful one. Disagreement remains about whether the disjunctive approach explains enough about the experience of deception.

### Related Concepts and Figures

- *The Myth of the Given* — experience and the demand for rational justification.
- *Direct Realism* — perceptual openness and world-involving evidence.
- *Indirect Realism and Sense Data* — alternatives to an inner intermediary.
- *Intentionalism* — debates over perceptual content and conceptual structure.
- *Wilfrid Sellars* — the space of reasons and criticism of the Given.
- *Ludwig Wittgenstein* — philosophical method and mind–world relations.
- *Immanuel Kant* — intuition, concepts, receptivity, and spontaneity.

### Further Reading


- McDowell, John. *Mind and World*. Harvard University Press, 1994; paperback with a new introduction, 1996.
- McDowell, John. *Mind, Value, and Reality*. Harvard University Press, 1998.
- McDowell, John. *Meaning, Knowledge, and Reality*. Harvard University Press, 1998.
- McDowell, John. *Having the World in View: Essays on Kant, Hegel, and Sellars*. Harvard University Press, 2009.
- McDowell, John. "Avoiding the Myth of the Given." In *Having the World in View*, 2009.

**Related Writing on This Site**

Mind, Matter, and Meaning presents the site author's arguments about mind and meaning. It is a separate argumentative work, not a neutral source for the account above.

**My View**

McDowell walks beside the book for a long stretch of the road, and I count him a fellow-traveler rather than an ally. The anti-Given lesson he shares with Sellars, that perception cannot serve knowledge as a tribunal of bare, unconceptualized data, enters the book's argument at several joints. His framing of second nature, openness to the world, and the space of reasons supplies vocabulary the book reaches for at structural moments.

I follow McDowell on the diagnosis of the inner-standpoint picture and on the anti-foundationalist methodology behind it. The transparency thesis in Tye's version and McDowell's openness-to-the-world thesis deploy the same broader insight in different ways: look for experience and you find the world.

Where I part ways: conceptualism. The book keeps Tye's nonconceptual grain. Perceptual experience carries a fineness of detail that outruns the concepts the perceiver commands, and animals and infants perceive without commanding concepts at all. "Nonconceptual" here names grain, not epistemic isolation, so declining conceptualism costs the anti-Given lesson nothing. Tye and Crane carve out the space for nonconceptual representational content, and I follow them. McDowell's quietism also sits too close to giving up on the constructive project the book pursues. Teleosemantics (in a supporting role) and causal-historical theories of reference answer the naturalization question McDowell prefers to dissolve. And on disjunctivism, the book credits the anti-veil motivation while declining the view. Treating worldly objects as constituents of experience makes misrepresentation in veridical perception difficult to explain, and intentionalist direct realism handles the same cases by a different mechanism. See *The Myth of the Given*, *Direct Realism*, and *Wilfrid Sellars*.

**Ruth Millikan**

Ruth Garrett Millikan's work in philosophy of mind and language develops a biological account of representation. Rather than beginning with an interpreter who assigns meanings to otherwise meaningless signs, she asks how a system's history and organization can establish what its signals are supposed to indicate. Her version of **teleosemantics**, often called **biosemantics**, gives particular importance to the mechanisms that use representations.

**Major Works and Development**

*Language, Thought, and Other Biological Categories* (1984) develops the framework of proper functions, intentional signs, and biological explanation. "Biosemantics" (1989) presents the account of representational content in a shorter form. *White Queen Psychology and Other Essays for Alice* (1993) extends the framework across questions about thought and psychology. *On Clear and Confused Ideas* (2000) applies it to concepts; *Beyond Concepts* (2017) develops the distinction between linguistic concepts and the individual cognitive capacities Millikan calls unicepts.

### Proper Functions

A proper function is not simply an effect a mechanism regularly produces. A heart pumps blood and makes noise, but the contribution to circulation explains why hearts of that kind were reproduced. Making noise does not. Millikan's account appeals to historically established families and the effects that explain the reproduction of their members.

This approach supplies a distinction between functioning and malfunctioning: a defective heart retains its function even when it cannot perform it. Proper functions may also be derived from the functions of mechanisms that produce or design something. Consequently, the framework does not restrict functions to traits directly selected by genetic evolution. Learning, reproduction of behavior, and cultural transmission require their own historical explanations, rather than the assumption that whatever currently proves useful already has a proper function.

### Producer–Consumer Organization

A signal's content depends on the organization linking its producer to its consumer. The consumer is a mechanism that uses the signal to perform a function. What matters especially is the correspondence between signal and world that helps explain how the consumer historically succeeded.

For example, a bee's dance varies with the location of a food source. Other bees use that variation to travel to the source. The dance's representational role is explained through a systematic mapping between its variations and locations, together with the functions of the mechanisms that produce and use it. Mere correlation with a location would not supply the whole account.

Millikan's **Normal conditions** are conditions in a historical explanation of successful functioning, not necessarily the statistically most common conditions. A correspondence can be Normal even when successful performance is infrequent. This qualification also prevents the theory from equating truth with success on every individual occasion: a false signal may accidentally lead to a useful result, while a true signal may fail because another part of the system malfunctions.

### Misrepresentation and Action

Biosemantics aims to explain how a representation can be false without relying on an interpreter's judgment. A signal can occur when the relevant world condition is absent, even though the consumer still responds as its organization disposes it to respond. Its present cause and its content therefore need not coincide.

Frog prey detection illustrates both the proposal and a difficulty. A mechanism may respond to small moving dark shapes, including inedible objects. Which feature fixes its content: flies, edible prey, or some broader condition relevant to feeding? Consumer function helps constrain the answer, but a history of successful feeding does not automatically distinguish every candidate description.

Millikan also develops **pushmi-pullyu representations**, which combine descriptive and directive functions. A bee's dance does not merely indicate where food is; it also guides where the receiving bees should go. Such cases challenge the assumption that representation must always divide neatly into belief-like description and desire-like instruction.

### Objections and Replies

**Content indeterminacy.** Critics ask whether selection history fixes sufficiently precise contents when several correlated features contributed to success. Millikan appeals to systematic mapping relations and detailed Normal explanations of consumer functioning, rather than fitness considered in isolation. Whether these resources uniquely determine sophisticated contents remains disputed.

**History and duplicates.** A spontaneously produced duplicate with no relevant history—the "Swampman" case—appears capable of thought, yet historical theories may initially deny it the same content as an ordinary person. Defenders distinguish present behavioral similarity from possession of historically constituted functions; critics question whether that distinction should govern mental attribution.

**Scope.** Explaining animal signals is not yet explaining mathematical thought, novel concepts, or the full flexibility of human language. Millikan extends the account through learning, language, and patterns of recognition. The adequacy of those extensions is a separate question from whether simple biological representations receive a plausible explanation.

Biosemantics primarily addresses content and misrepresentation. It does not, by itself, identify which content-bearing systems understand their representations or have conscious experience.

### Further Reading

- Millikan, Ruth Garrett. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. MIT Press, 1984.
- Millikan, Ruth Garrett. "Biosemantics." *The Journal of Philosophy* 86, no. 6 (1989): 281–297.
- Millikan, Ruth Garrett. *White Queen Psychology and Other Essays for Alice*. MIT Press, 1993.
- Schulte, Peter, and Karen Neander. "Teleological Theories of Mental Content." *Stanford Encyclopedia of Philosophy*.

### Related Concepts and Figures

- *Teleosemantics* — theories connecting content with biological function.
- *Reference* — how representations concern particular things or kinds.
- *Intentionality* — aboutness, accuracy, and misrepresentation.
- *Semantic Externalism* — historical and environmental contributions to content.
- *Fred Dretske* — a different function-based approach to naturalizing content.

### Related Writing on This Site

- Mind, Matter, and Meaning — the site author's philosophical argument, separate from this reference account.

### My View

Millikan supplies the principal teleosemantic framework behind the book's naturalization of content. Producer-consumer history, proper function, and a correspondence rule can explain how a state acquires accuracy conditions without an interpreter assigning them at the moment of use. The framework supports a qualified reference theory by showing how physical organization can become answerable to a world.

That route does not make every selected or trained mechanism a source of original content. A mechanism carries original content to the extent that its history and producer-consumer use fix accuracy conditions not exhausted by an interpreting practice. Fintan Mallory's word2vec analysis supplies a serious account of content about actual word-types and their context distributions. Whether that content is original at this limited grain remains unsettled. The vocabulary, corpus, task, and loss descend from public linguistic practice, but that ancestry alone does not establish semantic dependence. No further proprietor relation settles the issue.

Mechanism content also settles no larger attribution by itself. Integration, correction, memory, planning, and control bear on whether a whole system understands; perceptual mode, attention, and poising bear on phenomenal character and state consciousness. Millikan helps answer what a state represents and how it can misrepresent. The remaining questions require additional arguments from externalism, system organization, and the identity claim.

See *Teleosemantics*, *Reference*, and *Tyler Burge*.

### G. E. Moore

G. E. Moore (1873–1958) wrote on perception, knowledge, philosophical analysis, and ethics. His rejection of idealist arguments, distinction between awareness and its object, and defense of ordinary knowledge helped shape early analytic philosophy. These contributions do not constitute a single settled theory of perception: Moore continued to investigate sense-data and the relation between what is immediately experienced and physical objects.

#### Major Works

- "The Refutation of Idealism" (1903) — awareness, its objects, and an argument associated with *esse est percipi*.
- *Principia Ethica* (1903) — goodness, ethical analysis, and the open question argument.
- "The Status of Sense-Data" (1913–1914) — the relation between perceptual objects and physical things.
- "A Defence of Common Sense" (1925) — ordinary knowledge and philosophical accounts of it.
- "Proof of an External World" (1939) — a proof from the existence of hands to the existence of external objects.

#### Awareness and Its Object

"The Refutation of Idealism" examines the claim that being is being perceived. Moore argues that a sensation of blue includes two distinguishable elements: blue, the object of awareness, and awareness itself. Sensations of blue and green differ in their objects while sharing the element of consciousness. Identifying blue with the sensation of blue confuses the object with the awareness of it.

The argument attacks a particular route to idealism, not every possible idealist metaphysics. Distinguishing an object from awareness does not alone prove that the object can exist independently of every mind. Moore's title should therefore not be read as announcing a proof of all the claims ordinarily associated with realism.

### Diaphanous Experience

Moore observes that, in trying to introspect a sensation of blue, attention tends to fall on blue rather than on awareness. Awareness appears "diaphanous." Importantly, he also says that awareness can be distinguished if one looks attentively enough. His point supports an act–object distinction; it is not an unqualified claim that the mental act is impossible to attend to.

Later discussions of **transparency** draw on this passage. Representationalists argue that attending to experience reveals represented properties rather than additional intrinsic qualities of experience. Other theories explain the same observation differently, and Moore did not himself develop the later claim that phenomenal character is identical with representational content.

His subsequent engagement with sense-data complicates any portrayal of him as a straightforward contemporary direct realist. The question whether the immediately perceived color or shape belongs to an ordinary physical object remained a problem for his analysis.

### Goodness and the Open Question Argument

In *Principia Ethica*, Moore argues that goodness is simple and cannot be defined by identifying it with pleasure, desire, or another proposed property. Whatever definition is offered, it still seems meaningful to ask whether something possessing that property is good. Asking whether pleasure is good does not appear equivalent to asking whether pleasure is pleasure.

Moore calls attempts to define goodness in this way instances of the **naturalistic fallacy**, though his objection also concerns proposed metaphysical definitions, not just definitions using natural science. The term should not be conflated with the distinct problem of deriving an "ought" from an "is." His own account treats goodness as a non-natural property, while ethical judgment also requires inquiry into which things possess it.

A central objection distinguishes conceptual equivalence from property identity. A question can remain informative when two expressions refer to the same property, just as an empirical identity may be discovered without being obvious from the meanings of its terms. Ethical naturalists therefore deny that an open question alone rules out every naturalistic account. Moore's defenders reply that the persistent normative character of the question calls for explanation; the argument's force depends on what sort of reduction is proposed.

### Common Sense and the External World

Moore's 1939 proof begins with "Here is one hand" and "Here is another." If these are known premises, external objects exist. He distinguishes having a proof from being able to prove every premise of that proof, and does not claim to have independently eliminated every skeptical possibility.

Critics argue that the proof cannot persuade someone whose doubt concerns precisely whether the premises are known. Responses distinguish proving that external objects exist from supplying a non-question-begging answer to skepticism. Moore's broader approach gives ordinary knowledge greater initial credibility than philosophical premises that would undermine it; whether that priority amounts to a satisfactory epistemology remains debated.

**Further Reading**

- Moore, G. E. "The Refutation of Idealism." *Mind* 12, no. 48 (1903): 433–453.
- Moore, G. E. *Principia Ethica*. Cambridge University Press, 1903.
- Moore, G. E. "A Defence of Common Sense." In *Contemporary British Philosophy*, second series, edited by J. H. Muirhead, 1925.
- Moore, G. E. "Proof of an External World." *Proceedings of the British Academy* 25 (1939): 273–300.

**Related Concepts and Figures**

- *Transparency* — the later debate prompted by diaphanous awareness.
- *Indirect Realism and Sense Data* — Moore's continuing perceptual investigations.
- *Intentionalism* — a later interpretation of experiential transparency.
- *Bertrand Russell* — early analytic philosophy and theories of acquaintance.
- *Gilbert Harman* — a later formulation of the transparency argument.

**Related Writing on This Site**

- Mind, Matter, and Meaning — the site author's philosophical argument, separate from this reference account.

**My View**

Moore matters to this book primarily through the transparency observation, which I count among the three or four most important moves in the philosophy of perception. Harman made the datum central to the modern debate, and Tye developed it into a systematic argument for strong representationalism; Dretske supplies another major route to the explicit identity claim. The book treats transparency as powerful but limited evidence, not as a deduction: it counts against inner exhibits and supports the cumulative abductive case developed in Chapter 6. Moore saw the phenomenon and named it; Harman and Tye pressed its significance; the house identity claim belongs to the broader synthesis.

The refutation of idealism also serves the book as background, though not as a direct resource. Moore's insistence that the object of experience is distinct from the act of experiencing it steps in the right direction: it gestures toward the world-directedness of experience the book defends. Where Moore stops short is in the failure cases: hallucination, illusion, the bent stick in water. What is the object then? The book needs the Austinian demolition of the argument from illusion, plus the intentionalist account of perceptual error, to handle those cases; Moore's act/object distinction alone isn't enough.

The open question argument belongs to a different department of philosophy, and the book makes no direct use of it. But its structure has an analogue in the philosophy of mind. The hard problem functions similarly: whatever physical-functional story you tell, you can always coherently ask, "But why does it feel like anything?" Moore's argument that goodness resists definition in natural terms parallels Chalmers's argument that phenomenal consciousness resists reduction to physical facts. The book's response to Chalmers, that the explanatory gap is epistemic rather than ontological, is not available to Moore's metaethics. Spotting the structural parallel still helps. See *Transparency* and *Gilbert Harman*.

### Thomas Nagel

Thomas Nagel's philosophy examines the tension between a person's standpoint and an increasingly impersonal understanding of the world. His work on consciousness asks how an objective account can explain experience without losing its subjective character. His work on ethics, reason, and the natural order extends this question beyond the mind–body problem. The caution about reduction in his 1974 bat paper should be distinguished from the more ambitious anti-materialist arguments of *Mind and Cosmos* (2012).

### Major Works and Development

"What Is It Like to Be a Bat?" (1974) makes subjective character central to the problem of consciousness. *Mortal Questions* (1979) collects essays on death, moral luck, consciousness, and other limits of philosophical understanding. *The View from Nowhere* (1986) investigates the relation between subjective and objective perspectives across knowledge, action, and ethics. *Mind and Cosmos* (2012) argues that consciousness, cognition, and value require a revision of the materialist conception of nature.

### What It Is Like

Nagel characterizes a conscious organism as one for which there is something it is like to be that organism. This concerns experience, not merely the organism's behavior or its capacity to process information.

The bat example separates imagining unusual human behavior from understanding another creature's experience. Imagining oneself flying, hanging upside down, or detecting echoes gives a human imaginative reconstruction. It does not simply reveal what echolocation feels like to a bat. Differences in sensory organization can limit the concepts and imaginative resources available for describing another standpoint.

Nagel does not infer that bats' experiences are unreal, unknowable in every respect, or accessible only to one individual. A standpoint may be shared by creatures of a relevant kind. The difficulty is accounting for perspectival facts from an account designed to abstract away from particular standpoints.

### Reduction and Objective Phenomenology

Ordinary scientific reductions often separate how something appears to observers from what it is physically. Consciousness cannot be handled simply by relocating its appearances to another observer's mind: the appearances are part of the phenomenon requiring explanation.

The 1974 paper argues that existing reductive strategies do not adequately address this problem. It does **not** conclude that physicalism has been proved false. Nagel says that we lack a satisfactory conception of how it could be true, and proposes developing an **objective phenomenology**: a way of describing experiential structure that does not depend entirely on possessing the experience oneself.

Physicalist responses distinguish knowing all the physical facts from acquiring the capacities or acquaintance involved in undergoing an experience. On such accounts, the inability to imagine bat experience need not show that nonphysical facts exist. Nagel's challenge is broader than an

imaginative limitation: an explanation must make clear how the subjective character itself belongs to the natural account, rather than replacing it with behavioral or functional descriptions.

### The View from Nowhere

Objectivity develops by stepping back from features of one's particular perspective. Nagel regards this movement as a source of understanding, but not as a method that can simply eliminate subjectivity. A complete account must accommodate both the individual who acts and experiences and the impersonal world in which that individual appears.

The same tension arises in ethics: a person's projects and reasons matter from their own standpoint, while an impartial perspective may demand consideration of others. Nagel's project is to understand the relation between these perspectives, not to insist that objective inquiry is inherently futile.

### Mind and Cosmos

The later book argues that a reductive materialist account has difficulty explaining the existence of consciousness, the authority of reason, and objective value. Nagel distinguishes explaining how mental capacities are constituted from explaining their historical emergence. He considers broader monist possibilities and **natural teleology**: impersonal principles making the emergence of life and mind more intelligible or likely. He does not offer a completed teleological biology or defend a divine designer, and the book is not simply an endorsement of one established panpsychist theory.

### Criticism and Replies

Critics distinguish the extensive evidence for evolution from disputed metaphysical claims about reduction. H. Allen Orr challenges Nagel's assessment of evolutionary explanation; Elliott Sober questions the demand that an adequate explanation make remarkable outcomes antecedently probable. Evolution can produce capacities with uses extending beyond the circumstances in which they were selected, and an improbable event need not be inexplicable.

Nagel's position is that the issue concerns the intelligibility of consciousness, reason, and value within the natural order, not merely missing details in a biological history. Nevertheless, that philosophical concern does not itself establish new teleological laws. His positive alternative remains speculative, and disagreement about consciousness does not place evolutionary theory and an unspecified teleological replacement on equal evidential footing.

### Further Reading


- Nagel, Thomas. "What Is It Like to Be a Bat?" *The Philosophical Review* 83, no. 4 (1974): 435–450.
- Nagel, Thomas. *Mortal Questions*. Cambridge University Press, 1979.
- Nagel, Thomas. *The View from Nowhere*. Oxford University Press, 1986.
- Nagel, Thomas. *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature Is Almost Certainly False*. Oxford University Press, 2012.
- Orr, H. Allen. "Awaiting a New Darwin." *The New York Review of Books*, February 7, 2013.

**Related Concepts and Figures**

- *Qualia* — phenomenal qualities and what-it-is-like talk.
- *The Hard Problem of Consciousness* — later formulations of the explanatory challenge.
- *Physicalism* — physicalist accounts and their explanatory obligations.
- *Panpsychism* — one family of alternatives to reductive explanations.
- *David Chalmers* — a later development of the consciousness problem.

**Related Writing on This Site**

- Mind, Matter, and Meaning — the site author's philosophical argument, separate from this reference account.

**My View**

The bat paper serves the book throughout as an indispensable tool. "What is it like" points vividly at what needs explaining: the phenomenal character of experience, the felt quality of seeing red or tasting coffee. Nagel found the locution that makes the problem undeniable. Whatever else one thinks about his conclusions, his articulation of the target is exactly right.

The book parts ways with Nagel over what the subjective character of experience implies. The bat paper fixes the target with unmatched clarity, but it does not by itself establish that phenomenal facts fall outside the physical. The later Nagel increasingly treats the objective/subjective divide as evidence that the materialist framework cannot accommodate mind. My view locates the divide in our modes of access: description and acquaintance reach one fact by different routes.

*Mind and Cosmos* makes the stronger disagreement explicit. Nagel takes the explanatory difficulty as evidence that the standard materialist picture lacks something in nature and may need teleological principles. The book instead combines the identity claim with the acquaintance-recognition account. Phenomenal character consists in the complete world-presenting intentional profile; acquaintance explains why no description alone confers the first-person recognitional capacity. That leaves a real epistemic gap without adding teleological features to the universe.

Nagel sees what the problem is more clearly than almost anyone. Where I part from him is on what it implies. See *The Hard Problem of Consciousness* and *Physicalism*.

**Hilary Putnam**

Hilary Putnam (1926–2016) contributed to logic, philosophy of language, philosophy of mind, and debates about realism. His views changed substantially: early computational functionalism, semantic externalism, internal realism, and later natural realism address different problems and should not be collapsed into one doctrine. Several of his later arguments criticize positions he had previously helped develop, without ending debate over those positions.

**Major Works and Phases**

"Psychological Predicates" (1967), later reprinted as "The Nature of Mental States," develops a functional account of mental states. "The Meaning of 'Meaning'" (1975) presents Twin Earth and the linguistic division of labor. *Reason, Truth and History* (1981) develops internal realism. *Representation and Reality* (1988) criticizes computational accounts of intentionality. *The Threefold*

*Cord: Mind, Body, and World* (1999) rejects an interface conception of perception and develops natural realism.

### Functionalism and Multiple Realization

Early Putnam proposes that a mental state such as pain is a functional state of an organism rather than one particular physical–chemical state. Its role involves sensory inputs, relations to other internal states, and behavioral outputs. His model uses probabilistic automata, not just a deterministic input–output table.

The multiple-realization argument asks why pain in mammals, molluscs, and possible extraterrestrial organisms must always be the same physical kind. Functional organization promises a common description despite physical differences. This is a challenge to unrestricted type-identification of mental states with particular neural states, not a proof that every conceivable substrate can think. Nor does multiple realization by itself select functionalism over every alternative account.

### Twin Earth and the Division of Linguistic Labor

Twin Earth resembles Earth except that its superficially water-like liquid has a different underlying constitution, called XYZ. The example places speakers before the discovery of water's chemical composition, so that their ordinary beliefs and discriminatory abilities need not differ. Putnam argues that an Earth speaker's "water" concerns $H_2O$, while the counterpart's term concerns XYZ. What fixes reference is not exhausted by what either individual can discriminate internally.

A related argument concerns experts. An ordinary speaker may use "elm" and "beech" without being able to distinguish the trees. Membership in a linguistic community, including deference to knowledgeable speakers, can help fix the reference of those terms. **Semantic externalism** thus includes environmental and social dimensions. It is not the assertion that psychological states are irrelevant to meaning or that every aspect of meaning is determined externally.

### Realism and Its Revisions

Putnam's internal realism opposes the idea of one uniquely correct description of a ready-made world wholly independent of conceptual schemes. *Reason, Truth and History* explores the relation among description, truth, and idealized rational acceptability. This is not the simple view that people create physical objects by thinking about them.

His later natural realism rejects an additional barrier between mind and world. In *The Threefold Cord*, Putnam criticizes the **interface picture**, on which our cognitive contact is always first with inner items and only indirectly with external things. Successful perception can be perception of ordinary objects themselves. Neural mediation, illusion, and hallucination still require explanation; rejecting an inner epistemic screen does not mean denying physiological processing or error.

These phases involve revisions in his account of truth and perception. They should not be treated as interchangeable formulations of his 1975 externalism.

### The Critique of Computational Functionalism

In *Representation and Reality*, Putnam argues that computational organization does not uniquely individuate intentional states. Content can vary with environment and linguistic practice, while a shared intentional state may be realized through different computational arrangements. He also develops a triviality argument: under permissive mappings from physical states to machine states, implementation claims become too easy to satisfy to explain mentality.

That argument depends on the constraints governing implementation. It is not the uncontested result that every ordinary physical system implements every machine under any reasonable account. Defenders of computationalism require counterfactual support, appropriate causal organization, and independently specified mechanisms. Other functionalists widen the relevant roles to include environmental relations. Putnam's criticisms therefore distinguish machine-state functionalism from a broader claim that causal-functional organization matters to mind.

### Objections and Continuing Disputes

Twin Earth raises questions about which notions of content are externalist and whether internally individuated "narrow content" can still serve psychological explanation. Putnam's defenders distinguish the public reference of natural-kind terms from every cognitive difference between speakers.

Natural realism must explain how direct perceptual contact is compatible with indistinguishable hallucinations. Its response rejects the assumption that a common subjective appearance must imply a common inner object, but the positive account remains disputed.

Across these debates, Putnam's changes of position are part of the philosophical evidence to interpret, not a reason to declare either his early or late views automatically defeated.

### Further Reading

- Putnam, Hilary. "The Nature of Mental States." In *Mind, Language and Reality: Philosophical Papers*, vol. 2. Cambridge University Press, 1975. First published as "Psychological Predicates," 1967.
- Putnam, Hilary. "The Meaning of 'Meaning'." In *Mind, Language and Reality: Philosophical Papers*, vol. 2. Cambridge University Press, 1975.
- Putnam, Hilary. *Reason, Truth and History*. Cambridge University Press, 1981.
- Putnam, Hilary. *Representation and Reality*. MIT Press, 1988.
- Putnam, Hilary. *The Threefold Cord: Mind, Body, and World*. Columbia University Press, 1999.

### Related Concepts and Figures

- *Semantic Externalism* — environmental and social contributions to content.
- *Functionalism* — early functionalism and subsequent objections.
- *Reference* — natural-kind reference and linguistic deference.
- *Direct Realism* — perception without an epistemic interface.
- *Tyler Burge* — further development of social externalism.

**Related Writing on This Site**

- Mind, Matter, and Meaning — the site author's philosophical argument, separate from this reference account.

**My View**

Putnam stands as a fellow-traveler and a rival, depending on which turn of his career we follow. The book keeps his material anti-chauvinism, his externalism, and his rejection of the interface conception. It also takes seriously his demonstration that narrow functional and computational states classify contentful thought incorrectly in both directions. Meanings are not in the head, and internal role alone does not fix what a state concerns.

The externalist inheritance remains central. Reference reaches the world through environmental, causal-historical, and social relations rather than through a label attached to an inner concept. Kripke's chains, Millikan's producer-consumer history, and Burge's anti-individualism develop different parts of that result. The book's qualified reference theory therefore owes Putnam a substantial debt.

Two disagreements remain. Putnam considered a widened sociofunctionalism of organisms-cum-environments and ultimately rejected it because no finite equivalence rule could collect every realization of one belief. The book accepts the direction of the widening while declining the demand for finite algorithmic reduction: physically realized world-involving roles can remain objective and explanatory without forming a complete machine table. His natural realism also shares the anti-veil conclusion without supplying the house's positive intentionalist account. Putnam clears the ground and identifies the required width; the identity claim and teleosemantics provide further structure.

See *Semantic Externalism*, *Reference*, and *Functionalism*.

**Bertrand Russell**

Bertrand Russell (1872–1970) helped found analytic philosophy and transformed modern logic with Alfred North Whitehead. He wrote influentially on mathematics, knowledge, perception, language, mind, and physics, then carried philosophy into public life with unusual reach. Russell worked for much of his philosophical career in and around Cambridge and received the Nobel Prize in Literature in 1950. He also changed his mind often enough to leave scholars several Russells rather than one tidy system.

For *Mind, Matter, and Meaning*, the relevant Russell appears in his work on neutral monism, perception, and the limits of physical knowledge. He matters less for a completed theory of consciousness than for a question contemporary philosophy still has not put down. If physics tells us what matter does and how it relates, what—if anything—does it tell us about matter's concrete nature?

**Major Works**

- *The Principles of Mathematics* (1903) — logicism and the foundations of mathematics
- *Principia Mathematica* (with Alfred North Whitehead, 1910–1913) — the landmark formal development of logicism

- *The Problems of Philosophy* (1912) — acquaintance, description, sense-data, and the external world
- *Our Knowledge of the External World* (1914) — logical construction and scientific method in philosophy
- *The Analysis of Mind* (1921) — neutral-monist psychology, sensation, images, and belief
- *The Analysis of Matter* (1927) — physics, causal structure, perception, and the intrinsic-nature problem
- *Human Knowledge: Its Scope and Limits* (1948) — late epistemology and scientific inference

### Key Positions

**Knowledge by acquaintance and knowledge by description.** Russell distinguishes direct acquaintance with particulars or qualities from knowledge mediated by descriptions. The distinction shaped later discussions of Mary's Room, phenomenal concepts, and first-person knowledge, although contemporary acquaintance theories should not simply be read back into Russell unchanged.

**Neutral monism.** During the 1910s and 1920s Russell explored the idea that mind and matter can be constructed from events or particulars neutral between the mental and physical. An event counts as mental or physical through the systems of relations into which it enters rather than through membership in two fundamental substances. Russell's position changed across works, so "Russellian neutral monism" names a historical trajectory, not a single perfectly stable doctrine.

**Structural knowledge of the physical.** Here *structural* means knowledge of relations, causal roles, and mathematical organization. In *The Analysis of Matter*, Russell argues that physics can supply this knowledge while leaving open the complete character of what occupies the structure. Physical science tells us how events relate and what effects they produce. It may not tell us what the relata are apart from those roles.

**Percepts and the causal theory of perception.** Russell sought to connect the events known in perception with the causal order described by physics. Percepts belong to a subject's experiential history while also occupying causal relations within the physical world. Contemporary Russellian monists use this overlap as evidence that experience may disclose something about the concrete nature of physical events.

**Logical atomism and analysis.** Russell's broader method decomposes complex propositions and facts into simpler constituents and relations. Although logical atomism does not entail Russellian monism, its commitment to structural analysis helps explain why the relation between form and concrete occupant became so important in his philosophy of matter.

### From Russell to Russellian Monism

Contemporary Russellian monism begins from Russell's structural insight but develops a taxonomy he did not leave behind. Russellian panpsychism identifies fundamental *categorical* properties —what properties concretely are, not only what they do—with phenomenal properties. Panprotopsychism gives the fundamental base a nonphenomenal nature specially suited to constitute consciousness. Panqualityism posits qualities without microsubjects. Powerful-quality versions identify causal power and qualitative character as one property.

These positions should not all be attributed to Russell. The modern family grew from his question, then acquired metaphysical machinery from later debates about physicalism, phenomenal concepts, combination, and causation. Russell supplied the opening; Chalmers, Strawson, Goff, Heil, and others built different houses on it.

### My View

Russell's structural warning deserves a permanent place in the book. Physics does not become a complete vocabulary for lived reality merely by becoming a complete causal science. Description does not confer acquaintance, and equations do not substitute for seeing red or feeling pain.

The warning underdetermines the metaphysics. Physics's structural character does not prove that distinct categorical realizers exist; pure-power and structuralist views remain live. Even if a categorical ground exists, physics does not select a phenomenal rather than nonphenomenal occupant. Acquaintance with an organism-level phenomenal profile does not establish that fundamental matter has a qualitatively continuous nature.

The missing step is the inheritance premise: the assumption that a nonrelational macroquality must derive from qualitative microproperties rather than consist in relational organization at the level of the whole. Russell's work gives us no settled argument for that premise.

The book draws a local lesson from Russell's insight. Acquaintance may disclose something of the concrete physical episode while leaving its fundamental realizers unknown. Phenomenal character can count as categorical in the harmless sense that the shape of a whole does: concrete and real, although relations constitute it. If this only redescribes the realized relational profile, the local option converges with the book's type-B physicalism. A necessarily co-instantiated categorical realizer remains metaphysically possible, but needs independent support or explanatory work; a stable matched-profile contrast would defeat the identity, while its absence supports economy without proving equivalence.

Russell therefore matters to the project as a diagnostician of physical description's limits, not as an authority for panpsychism. His insight keeps the physicalist modest. It does not make consciousness fundamental.

## John Searle

John Searle's work spans speech acts, intentionality, consciousness, and social institutions. In philosophy of mind, he argues that executing a computer program is not sufficient for understanding and that consciousness must be explained through the causal powers of physical systems. His positive account, **biological naturalism**, rejects both substance dualism and theories that eliminate first-person experience.

### Major Works and Development

*Speech Acts* (1969) investigates linguistic actions and their rules. "Minds, Brains, and Programs" (1980) introduces the Chinese Room. *Intentionality* (1983) develops an account of mental directedness. "Is the Brain a Digital Computer?" (1990) questions computational descriptions of nature. *The Rediscovery of the Mind* (1992) develops biological naturalism. *The Construction of Social Reality* (1995) extends his account of collective intentionality to institutions.

### The Chinese Room

A person who does not understand Chinese receives Chinese symbols and follows a rulebook for returning appropriate symbols. The instructions can enable convincing answers without giving the operator an understanding of Chinese. Searle uses the case against **strong AI**: the claim that an appropriately programmed computer thereby understands and that its program explains that understanding. He distinguishes this from using computers as useful models or tools for studying cognition.

The argument turns on the relation between formal processing and meaning. A program specifies operations over syntactic features. Searle argues that those operations do not, by themselves, supply semantic understanding. The disputed step is from the operator's lack of understanding to the conclusion that the implemented system does not understand either.

### Replies and Searle's Responses

**The systems reply** locates understanding in the room's whole organization, including rules and memory, not in the operator considered alone. Searle responds by imagining internalizing the entire organization: he could memorize the rules and perform all operations while still understanding no Chinese. Critics maintain that ignorance at the operator level does not determine what capacities the organized system realizes, even when one person implements it.

**The robot reply** adds perceptual and motor contact with the world. Searle argues that adding signals and movements to formally specified processing need not give their meaning to the operator. Defenders reply that the relevant claim concerns the embodied system's learning and world-related organization, not the operator's awareness of each computation.

**The brain-simulator reply** proposes reproducing the relevant organization of a Chinese speaker's brain. Searle distinguishes simulating its formal pattern from duplicating the causal powers that produce understanding. Critics ask whether a sufficiently faithful realization of that organization already duplicates the relevant powers. Calling it a simulation cannot settle that question.

These replies concern sufficiency and implementation. They do not establish that fluent behavior never supplies evidence of understanding, nor does the Room conclusively demonstrate that every artificial system must lack it.

### Original and Derived Intentionality

Searle distinguishes the intentionality of mental states from that of words, maps, and other artifacts. An inscription means something through its use by minded agents; a belief or intention has aboutness in its own right. This is his distinction between original or intrinsic intentionality and derived intentionality.

He also emphasizes the **Background** of capacities and the **Network** of other intentional states that make particular beliefs and intentions intelligible. Understanding is not merely attaching an interpretation to isolated symbols. Opponents dispute whether original intentionality requires Searle's account of consciousness or whether it can be explained through causal, functional, or historical relations.

### Biological Naturalism

On Searle's account, lower-level neurobiological processes cause and realize consciousness as a higher-level feature of a brain system. Consciousness has a first-person ontology: being experienced is part of what it is. He therefore distinguishes causal explanation from eliminating or ontologically reducing its subjective character.

The view does not posit a separate immaterial substance. It also does not demonstrate that only carbon-based systems could be conscious. An artificial system would need the relevant causal powers; a program description alone would not establish that it has them. Searle's analogies with simulated digestion or weather illustrate his objection, but critics question whether the same implementation distinction applies to understanding.

### Computation and Further Objections

Searle's later argument treats syntax and computational descriptions as observer-relative. Critics distinguish arbitrary mappings from computational organization constrained by a mechanism's causal and counterfactual structure. They argue that rejecting unconstrained interpretation need not make every physical computation merely assigned.

Biological naturalism also faces questions about mental causation: how can an ontologically irreducible conscious feature be causally effective without competing with its physical realization? Searle's reply treats the conscious and neurobiological descriptions as levels of one system, not independent competing causes. Whether this coherently combines irreducibility and causal unity remains disputed.

### Further Reading


- Searle, John R. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, 1969.
- Searle, John R. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3, no. 3 (1980): 417–424, followed by commentary and response.
- Searle, John R. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press, 1983.
- Searle, John R. "Is the Brain a Digital Computer?" *Proceedings and Addresses of the American Philosophical Association* 64, no. 3 (1990): 21–37.
- Searle, John R. *The Rediscovery of the Mind*. MIT Press, 1992.


### Related Concepts and Figures

- *The Chinese Room* — the thought experiment and its replies.
- *Intentionality* — original and derived aboutness.
- *Simulation and Realization* — models, implementations, and causal powers.
- *Functionalism* — a principal target and source of responses.
- *Alan Turing* — behavioral approaches to machine intelligence.

### Related Writing on This Site

- Mind, Matter, and Meaning — the site author's philosophical argument, separate from this reference account.

### My View

Searle supplies several of the book's sharpest instruments, but each now carries a bounded assignment. The Chinese Room establishes that operator-level formal rule-following and fluent output do not establish understanding; it does not defeat an implemented whole that genuinely realizes content-bearing, integrated cognition. His observer-relativity argument succeeds against unconstrained implementation mappings, while the book grants that disciplined, intervention-supporting causal-computational organization can belong intrinsically to a system. And the simulation/realization distinction poses a real question about causal powers without letting the word *simulation* decide whether a concrete implementation realizes a mind.

The deepest departure concerns biological naturalism and the hard problem. Searle treats consciousness as caused by and realized in neurobiology while retaining its first-person ontology and ontological irreducibility. The book instead adopts a Type-B identity: phenomenal character consists in the complete world-presenting intentional profile of a conscious state. Correlation or cross-level causation cannot retire the explanatory bridge; identity can. Biological naturalism can remain wholly naturalistic and anti-dualist, but it structurally cannot make the identity-based deflation on which the book depends.

No biological substrate receives a monopoly in advance. An artificial or virtual system could qualify if it actually realized the relevant content-fixing history, producer-consumer organization, integration, perceptual mode, and poising. That is not generic multiple realization by program identity; recurrence has to be shown capacity by capacity. The book therefore keeps Searle's bounded diagnostics, rejects both biological exclusivity and abstract-program sufficiency, and develops its positive account through representational identity, teleosemantics, and world-involving organization.

### Wilfrid Sellars

Wilfrid Sellars (1912–1989) worked on knowledge, language, perception, and the relation between ordinary understanding and science. His critique of the **Myth of the Given** challenges the idea that empirical knowledge rests on sensory items whose authority is independent of conceptual capacities. He also develops a positive account of thought and sensation, rather than denying that inner episodes exist.

### Major Works

"Empiricism and the Philosophy of Mind" (1956) examines perceptual knowledge and reconstructs the language of inner episodes. "Philosophy and the Scientific Image of Man" (1962) distinguishes the manifest and scientific images. *Science, Perception and Reality* (1963) collects these and related essays. *Science and Metaphysics: Variations on Kantian Themes* (1968) develops a more systematic account of perception, conceptual structure, and scientific reality.

### The Myth of the Given

Sellars distinguishes a sensation's causal role from its capacity to justify a claim. A sensory occurrence can cause someone to respond without thereby constituting knowledge. Knowing that something is red requires more than a reliable disposition to utter "red": it involves a place within practices of recognizing circumstances, making claims, and responding to reasons.

The problematic Given is supposed to supply authoritative knowledge while being independent of the conceptual capacities needed for knowledge. Sellars argues that this combination fails. If a sensory item has no role in a justificatory framework, its occurrence alone cannot provide the reason that the foundationalist requires.

This does not mean that every perceptual belief is inferred from other beliefs. Observation reports may be noninferential while depending on acquired conceptual abilities and background competence. Nor is the critique simply a denial of nonconceptual sensation. Sellars distinguishes sensory episodes from the epistemic authority sometimes attributed to them.

### Looks, Sees, and the Space of Reasons

Saying that someone **sees that** a tomato is red normally endorses the truth of the report. Saying that it **looks** red allows that endorsement to be withheld. Sellars uses this contrast to investigate perceptual appearance without assuming that a red inner object must be inspected whenever something looks red.

His **logical space of reasons** concerns justification, inference, and rational responsibility. A causal explanation of why someone makes a report differs from an account of why the report counts as knowledge. This distinction is not automatically a division into two substances or two independent worlds; it marks different explanatory and normative roles.

### The Myth of Jones

Sellars imagines a community that can describe public behavior and speech but lacks a vocabulary for inner episodes. Jones introduces thoughts as theoretical states modeled on meaningful overt utterances. Members learn to use this vocabulary to explain one another and subsequently to make first-person reports.

A further stage introduces impressions to account for perceptual similarities and differences, including cases in which things merely seem a certain way. Sellars distinguishes the theory's modeled structure from the mistaken assumption that its inner episodes literally possess every property of their models.

The story is a philosophical reconstruction of how concepts of inner episodes could acquire a role, not an empirical history of prehistoric language. It neither makes thought identical with audible speech nor makes people continuously infer their own thoughts from behavior. A concept can originate in explanatory practice and later support direct, noninferential self-report.

### Manifest and Scientific Images

The **manifest image** includes the framework in which people understand themselves as persons who act for reasons, possess intentions, and participate in norms. It is not merely the momentary appearance of things or a collection of childish errors. The **scientific image** develops explanatory accounts that posit entities and processes beyond ordinary observation.

Sellars seeks a synoptic understanding of their relation. He gives science authority in explaining what exists, while treating personhood and normative discourse as indispensable to understand-

ing ourselves. Reconciling sensory qualities and normative life with the scientific account remains a task, not a result supplied just by naming the two images.

**Objections and Developments**

Critics ask whether conceptual requirements make perceptual knowledge too demanding for children and animals, and whether a causal account can underwrite justification without requiring the subject to articulate every supporting reason. Sellars's distinction between noninferential reporting and conceptual dependence helps frame the reply, but leaves disputes about how much competence is necessary.

Another objection asks how experience constrains judgment if sensations are excluded from the space of reasons. John McDowell emphasizes the conceptual character of experience; Robert Brandom develops the normative and inferential dimensions of content. Their projects draw on Sellars without constituting one uniform "Sellarsian" position.

Debate also concerns the status of sensory qualities in his scientific account. Rejecting epistemically self-authenticating sensory foundations does not by itself resolve the metaphysics of consciousness.

**Further Reading**

- Sellars, Wilfrid. "Empiricism and the Philosophy of Mind." In *Minnesota Studies in the Philosophy of Science*, vol. 1, edited by Herbert Feigl and Michael Scriven, 253–329. University of Minnesota Press, 1956.
- Sellars, Wilfrid. "Philosophy and the Scientific Image of Man." In *Frontiers of Science and Philosophy*, edited by Robert G. Colodny. University of Pittsburgh Press, 1962.
- Sellars, Wilfrid. *Science, Perception and Reality*. Routledge & Kegan Paul, 1963.
- Sellars, Wilfrid. *Science and Metaphysics: Variations on Kantian Themes*. Routledge & Kegan Paul, 1968.

**Related Concepts and Figures**

- *The Myth of the Given* — sensory causation and epistemic authority.
- *Inner Speech* — speech, thought, and first-person reporting.
- *Phenomenology* — experience as a subject of description and explanation.
- *John McDowell* — a later account of experience and reasons.
- *Ludwig Wittgenstein* — psychological language and public practices.

**Related Writing on This Site**

- Mind, Matter, and Meaning — the site author's philosophical argument, separate from this reference account.

**My View**

Sellars stands under more of this book than the prose shows: less explicit than Searle or Tye, more pervasive in methodology. The structural diagnosis of the Myth of the Given, the Jonesean account of inner-episode talk, and the broader anti-foundationalist approach are all Sellarsian in origin.

I deploy Sellars's anti-Given diagnosis against three specific targets. Sense-datum theory assigns immediately apprehended inner items the task of grounding empirical knowledge, and they cannot do both jobs. Some foundationalist qualia theories cast qualia in the same role, as intrinsic, immediately accessible items that ground knowledge, and Sellars's diagnosis applies there. A sophisticated mental-paint theorist can instead treat qualia as intrinsic properties of experiential states without making them perceived inner objects or infallible evidence; that stronger view requires Chapter 6′s separate identity argument. The Cartesian inner standpoint, where introspection supposedly delivers self-authenticating reports about inner items, remains the third target.

I also extend the Jonesean myth to inner speech specifically: inner-voice talk is theoretical vocabulary modeled on overt verbal utterance, not a window onto pre-given private contents. The inner voice is public speech gone quiet.

Where the book is more cautious than full Sellarsianism requires: I do not fully embrace inferentialism, whose Brandomian development is more apparatus than the main text needs. And I lean toward causal-historical accounts of reference (Putnam, Burge, Kripke) over functional-role semantics. The book is Sellarsian where Sellars's structural arguments do work for it; it stops short of the technical positive program. See *The Myth of the Given*, *Inner Speech*, and *John McDowell*.

**Anil Seth**

Anil Seth studies consciousness through neuroscience, computational modeling, and experiments on perception and self-experience. His work connects **predictive processing** with the body's internal regulation. The phrase **controlled hallucination** summarizes his account of perceptual construction; the **beast machine** proposal connects conscious selfhood with the processes that sustain a living organism.

**Major Works and Research Themes**

"Interoceptive Inference, Emotion, and the Embodied Self" (2013) proposes a predictive account of bodily feeling and self-experience. The 2017 TED talk "Your Brain Hallucinates Your Conscious Reality" presents the framework for a broad audience. *Being You: A New Science of Consciousness* (2021) integrates perception, measures of conscious level, and the biological basis of selfhood.

"Conscious artificial intelligence and biological naturalism," first published online in 2025, develops his case that consciousness depends on living organization and questions whether computation alone suffices. It considers what more brain-like or life-like artificial systems might require.

These themes involve distinct questions: whether someone is conscious, which contents they experience, and how they experience themselves. A theory of perceptual content is not automatically a full account of conscious level or subjecthood.

**Perception as Predictive Inference**

Sensory signals underdetermine their external causes. A similar pattern at the retina can arise from different combinations of illumination, surface properties, and viewpoint. Predictive

processing explains perception through interaction between expectations about those causes and sensory evidence that corrects them.

In predictive-coding models, generative predictions interact with prediction errors across processing levels. The relative weight assigned to signals depends on their estimated reliability, often called **precision**. "Inference" here need not involve conscious reasoning or an inner observer: it describes a proposed computational organization of perceptual processing.

An ambiguous sound becoming clearer after one learns what was said illustrates how prior information can change experience. Such demonstrations motivate the role of expectations, but do not uniquely establish every detail of a predictive-coding architecture.

### Controlled Hallucination

Seth uses the phrase to emphasize that perception depends on active prediction rather than passively copying sensory input. His stronger claim identifies conscious perceptual contents with the joint content of top-down perceptual predictions: the brain's best guesses about the causes of sensory signals. This goes beyond saying that expectations sometimes influence perception. In ordinary perception, sensory contact constrains the predictions. Hallucinations involve different relationships between prediction and constraint.

The slogan does not mean that every perceived object is nonexistent or that successful perception and clinical hallucination are identical in every respect. A prediction can concern a real external object and be accurate. However, calling all perception hallucination can obscure those differences, and philosophical critics question what follows from the constructive mechanism to the nature of perceptual contact.

Predictive processing itself does not logically require a mental screen standing between the subject and the world. Representational, enactive, and relational interpretations disagree about that issue while accepting various roles for prediction. Seth's framing is one account within this debate, not a consensus metaphysical conclusion of neuroscience.

### Interoception and the Beast Machine

**Interoception** concerns the body's internal physiological condition. Seth proposes that emotions and aspects of embodied selfhood depend on predictive models of internal signals, integrated with vision, touch, proprioception, and other information.

The rubber-hand illusion provides an example of bodily self-experience's flexibility. When a visible artificial hand and a hidden real hand receive suitably coordinated stimulation, a participant can experience the artificial hand as belonging to their body. Interoceptive and multisensory approaches investigate how reliability, expectation, and bodily signals affect this attribution.

The beast-machine account emphasizes regulation, not just representation of the body as an external object. Keeping physiological variables within viable bounds can explain why bodily feeling has an affective character and why self-experience is tied to an organism's continued existence.

In his later **biological naturalism**, Seth argues that consciousness depends on our nature as living organisms, rather than on computation alone. He judges real artificial consciousness unlikely along current AI trajectories, while regarding more brain-like or life-like systems as more plausible candidates. These are his substantive claims, not experimentally established necessities. They do not amount to a restriction to carbon as such, and critics can challenge whether biological organization is required or merely one route to consciousness.

### The Real Problem of Consciousness

Seth proposes explaining why particular experiences have the properties they do by building systematic links between phenomenology and mechanisms. He calls this the **real problem**, distinguishing it from both cataloguing functional capacities and answering why physical processes produce experience at all.

The strategy allows empirical progress without first settling every metaphysical question. Critics maintain that explaining discriminations, reports, or even experiential structure may leave the hard problem untouched. Seth's response is methodological: detailed explanatory bridges may change what remains puzzling, rather than merely postpone investigation until philosophy is settled.

### Evidence, Objections, and Open Questions

Expectations, bodily signals, and sensory reliability demonstrably affect perception. What remains disputed is how broadly predictive architectures explain those effects and what distinctive predictions separate them from competing models. Flexibility becomes a weakness if a model can accommodate every outcome only by changing its assumptions after the fact.

Questions also remain about which predictive processes are conscious, how conscious level relates to content, and whether bodily regulation is necessary for consciousness or especially important for biological selfhood. Seth's account supplies hypotheses for investigating these questions; it should not be presented as either a complete solution or merely a provocative slogan.

### Further Reading

- Seth, Anil K. "Interoceptive Inference, Emotion, and the Embodied Self." *Trends in Cognitive Sciences* 17, no. 11 (2013): 565–573.
- Seth, Anil. "Your Brain Hallucinates Your Conscious Reality." TED talk, 2017.
- Seth, Anil. *Being You: A New Science of Consciousness*. Faber & Faber, 2021.
- Seth, Anil K. "Conscious artificial intelligence and biological naturalism." *Behavioral and Brain Sciences*. First published online April 21, 2025. doi:10.1017/S0140525X25000032.

### Related Concepts and Figures

- *Predictive Processing and the Interface* — predictive mechanisms and their philosophical interpretations.
- *Direct Realism* — perceptual contact with external objects.
- *Embodiment* — bodily organization and experience.
- *Vehicle and Content* — representational processes and what they represent.
- *Giulio Tononi* — a different approach to consciousness and causal organization.

**Related Writing on This Site**

- Mind, Matter, and Meaning — the site author's philosophical argument, separate from this reference account.

**My View**

Predictive processing competes directly with the book's spine. If what you are aware of is the brain's internal model, then experience presents you with a representation, not the world: the inner screen, rebuilt in the vocabulary of Bayesian inference. Whether the framework actually entails that conclusion, or merely seems to, is one of the book's central quarrels.

I take Seth seriously because he is mostly right about the *machinery* and, I think, wrong about what the machinery shows. The brain predicts, models, and corrects; perception runs as an active, inferential, top-down affair. All of that I grant without reservation, and the science is excellent. The trouble is the slide from a true claim about sub-personal mechanism to a false claim about the object of awareness: from "the brain infers the world" to "what you perceive is a hallucination." That is the vehicle/content confusion at the neural level. The predictive machinery is the *vehicle*; it does not follow that the machinery is what you are aware *of*. When the inference runs well, what it delivers to you is the world — the tomato, not a tomato-shaped guess hung behind your eyes.

Call experience a "hallucination" and you have already conceded the inner theater the book spends its first part tearing down — you have just repainted the screen in Bayesian colors. The better description: perception is the brain's way of *putting us in contact with* our surroundings, fallibly and actively, not a way of trapping us inside a model of them. Here Seth and I share more than the slogan suggests. His attention to interoception and regulation may illuminate autonomy, individuality, vulnerability, and welfare. But the book no longer treats a biological body, reciprocal self-maintenance, or a practical stake as the floor of genuine aboutness. Content may be fixed through selection, learning, producer-consumer use, and world-involving history, including in artificial mechanisms; subjecthood and consciousness require further organization. Our central dispute remains about the object of perception: the brain predicts, but when perception succeeds it reaches the world rather than presenting an inner hallucination.

**Claude Shannon**

Claude Shannon (1916–2001) developed mathematical tools for switching circuits and communication. His information theory quantifies uncertainty, coding, and transmission reliability. It deliberately separates the engineering question of reproducing a message from questions about what that message means. This makes the theory widely applicable, while limiting what its quantities alone establish about representation or understanding.

**Major Works and Dates**

"A Symbolic Analysis of Relay and Switching Circuits" originated as Shannon's MIT master's thesis in 1937 and appeared as a journal article in 1938. "A Mathematical Theory of Communication" appeared in two parts in 1948. The 1949 book *The Mathematical Theory of Communication* combines Shannon's paper with an introductory essay by Warren Weaver. The article's "A" and the book's "The" belong to different publication titles.

### Logic and Physical Circuits

Shannon showed how Boolean algebra could be used to analyze and synthesize relay and switching circuits. Connections in series and parallel can implement logical conditions, allowing circuit behavior to be specified and simplified symbolically.

This is a bridge between an abstract logical description and an engineered physical mechanism. It does not depend on a particular interpretation of the circuit's signals as English sentences. The contribution concerns implementation of logical operations, not a conclusion that every system implementing those operations thereby understands their semantic use.

### Entropy and Information

For a discrete source with outcome probabilities $p_i$, Shannon entropy is:

$H(X) = -\Sigma_i \, p_i \log_2 p_i$.

Using base-two logarithms expresses entropy in bits. The quantity measures average uncertainty about which outcome will occur. The information associated with a particular outcome—the outcome's surprisal—is $-\log_2 p_i$. Entropy averages that quantity across possible outcomes; the two should not be conflated.

A predictable symbol carries little new information in this sense. A less probable symbol is more surprising. Neither fact makes the symbol more meaningful, more important, or more truthful. A source of random symbols can have high entropy without expressing an intelligible message.

Shannon credits John W. Tukey for the word "bit." Information theory made the unit central; it did not require claiming that Shannon invented the word or every earlier idea about communication.

### Channel Capacity and Coding

A communication system involves a source, transmitter, channel, receiver, and destination, with noise potentially altering the signal. The statistical properties of the source determine opportunities for compression. The channel's properties limit the rate of reliable transmission.

The noisy-channel coding result establishes that, under specified assumptions and with sufficiently long suitable codes, transmission below capacity can make error probability arbitrarily small. It does not promise perfect communication at any finite block length or specify one universally practical code. The distinction between a mathematical existence result and an efficient engineering implementation is important.

**Mutual information** concerns the statistical dependence between source and received variables: how much knowing one reduces uncertainty about the other. Channel capacity optimizes the achievable informational rate over appropriate input distributions. These quantities depend on a specified probabilistic model, not on semantic interpretation alone.

### Why Meaning Is Set Aside

Shannon explicitly says that semantic aspects are irrelevant to the engineering problem he defines. The system must reproduce a selection from possible messages, whether the selection expresses a true statement, a false statement, or no statement at all.

This is a scope restriction, not a denial that messages have meanings. Nor does it imply that statistics can never contribute to explanations of representation. Later theories attempt to connect informational relations with learning, biological function, or linguistic practice. Those additional proposals must explain reference and misrepresentation; the Shannon equations do not already contain them.

**Applications and Limits**

The framework can be applied to electronic communication, neural signals, and other physical processes when their states and distributions can be modeled appropriately. Its generality does not mean that all applications concern the same phenomenon or that information is a substance moving through every system.

For neural data, an informational dependence may identify a useful relation between activity and stimuli without establishing what the organism consciously experiences. For language, reducing prediction error may measure a statistical achievement without directly measuring truth or understanding. Conversely, a system can transmit a meaningful message using an informationally ordinary channel. Statistical information and semantic content are distinct questions, not mutually exclusive kinds of system.

**Further Reading**

- Shannon, Claude E. "A Symbolic Analysis of Relay and Switching Circuits." *Transactions of the American Institute of Electrical Engineers* 57, no. 12 (1938): 713–723. Based on the 1937 MIT master's thesis.
- Shannon, Claude E. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27 (1948): 379–423 and 623–656.
- Shannon, Claude E., and Warren Weaver. *The Mathematical Theory of Communication.* University of Illinois Press, 1949.

**Related Concepts and Figures**

- *Information and Meaning* — statistical dependence and semantic content.
- *The Symbol Grounding Problem* — how symbols acquire meaningful use.
- *Teleosemantics* — biological-function approaches to representation.
- *Fred Dretske* — an attempt to connect information with content.
- *Alan Turing* — the formal theory of computation.

**Related Writing on This Site**

- Mind, Matter, and Meaning — the site author's philosophical argument, separate from this reference account.

**My View**

Shannon's distinction between signal and meaning is the hinge of an entire chapter of this book. A system can carry, store, and transform information flawlessly while meaning nothing by it. That is not a failure of engineering but the explicit boundary of what information theory measures. Every later attempt to read semantics off statistics, from early symbol-processing AI to today's large language models, runs up against the line Shannon drew on his first page.

Shannon is an ally I lean on precisely because he refused to overreach. He gave us the cleanest available statement of a distinction the book's middle chapters turn on. Carrying information is one thing, meaning something by it is another, and no amount of the first adds up to the second. A radio carries the broadcast without understanding a syllable; a photograph carries information about the scene without being *about* it in the way a thought is. Shannon told us how to measure the carrying. He was scrupulous enough to say, out loud, that the measuring left meaning untouched.

The mischief came later, from readers who took information theory's reach as a theory of mind. The book keeps Shannon's measure and semantic content distinct without treating every information claim as metaphor. Shannon information is a real statistical relation; representational content adds correctness conditions fixed through selection, learning, producer-consumer use, and world-involving history. Those conditions may arise in biological or artificial mechanisms. They do not by themselves settle whole-system understanding, consciousness, or autonomous stake. Bits can carry and constrain a representation, but bits alone do not determine what it is about.

### Giulio Tononi

Giulio Tononi develops **Integrated Information Theory** (IIT), a theory connecting consciousness with a system's intrinsic causal organization. The proposal begins with properties attributed to experience and asks what physical systems could satisfy corresponding requirements. Its successive versions differ in their mathematical definitions. The account developed with Larissa Albantakis and colleagues as IIT 4.0 should not be reduced to the slogan that consciousness is one numerical "information score."

### Major Works and Development

"An Information Integration Theory of Consciousness" (2004) formulates an early account using differentiation and integration. *Phi: A Voyage from the Brain to the Soul* (2012) presents the ideas through a popular narrative. Later collaborative papers develop IIT 3.0 and 4.0. Albantakis and colleagues' 2023 paper, "Integrated Information Theory (IIT) 4.0," specifies revised measures and an explicit account of causal distinctions and relations.

### From Experience to Physical Postulates

IIT treats the existence of experience as its starting point. It then identifies five proposed essential properties: **intrinsicality**, **information** or specificity, **integration**, **exclusion** or definiteness, and **composition** or structure. An experience exists for the experiencer, is the particular experience it is, forms a unified whole with definite boundaries, and contains differentiated aspects related within that whole.

The theory maps these claims onto physical postulates. A candidate substrate must possess cause–effect power for itself, constrain particular causes and effects, do so irreducibly as a whole, have a definite boundary and grain, and support a structured set of causal contributions. The mapping is an explanatory proposal. The phenomenological axioms alone do not constitute a mathematical proof that precisely these physical postulates follow.

### IIT 4.0 and Phi

IIT analyzes counterfactual causal constraints: how possible interventions and states would alter the system's possible causes and effects. This is not ordinary Shannon information about an external message, and it is not simply the amount of observed correlation among recorded units.

In IIT 4.0, system integrated information, denoted $\varphi_s$, helps identify a maximally irreducible substrate or **complex** among overlapping candidates. The complex specifies a **Φ-structure** of causal distinctions and relations. The structure's integrated information Φ concerns those contributions; it is not interchangeable with $\varphi_s$ or with every earlier version's use of phi.

The theory identifies an experience with the cause–effect structure unfolded from its maximal substrate. The proposed explanation concerns both the existence and the particular quality of experience. A single scalar cannot describe the full structure it posits.

### Consciousness and Behavior

IIT separates intrinsic causal organization from input–output competence. Two systems could produce matching outputs while differing in their intrinsic organization and, according to IIT, their experience. A strictly feed-forward multiunit system lacks the irreducible whole required for consciousness at that system level, even if it computes a sophisticated function. This need not mean that every component within any feed-forward arrangement has no causal powers or could not qualify separately under the theory.

Conversely, a relatively simple system can qualify as conscious if it has the required intrinsic irreducibility. The theory therefore allows consciousness beyond familiar biological organisms. It does not assign consciousness indiscriminately to every collection of matter: exclusion and maximality govern which subsets and grains count as subjects. Whether these principles supply a convincing account of subject boundaries remains disputed.

### Empirical Work and Measurement Limits

IIT motivates investigations of sleep, anesthesia, brain injury, cortical organization, and responses to perturbation. Measures of complexity and responsiveness may be useful clinically even when they are not exact calculations of IIT's quantities.

Calculating the complete structure requires a causal model and extensive evaluation of candidate mechanisms, partitions, and grains. This is difficult for large biological systems. Evidence for a practical complexity measure should therefore not automatically be presented as confirmation of IIT's full metaphysical identity claim.

### Criticisms and Replies

**The axioms–postulates step.** Critics question whether introspection warrants the proposed axioms and whether they uniquely support IIT's formalism. Proponents present the framework as an inference to a good explanation whose detailed consequences can be investigated.

**The unfolding argument.** Doerig and colleagues argue that matching functional behavior can be implemented with different causal structures, challenging how causal-structure theories can be tested through behavioral evidence. IIT defenders reject behavior as a constitutive criterion and

dispute the argument's assumptions about theory-independent evidence. The dispute concerns testing and inference, not an agreed experimental disproof or an agreed refutation of the objection.

**Unfamiliar subjects and contents.** Critics challenge consciousness assignments to simple systems and ask how intrinsic structure explains familiar experiential qualities. Defenders argue that unfamiliar consequences need not refute a principled theory and seek structural correspondences between physical organization and phenomenology. Demonstrating those correspondences, rather than merely asserting them, remains a central research obligation.

### Further Reading

- Tononi, Giulio. "An Information Integration Theory of Consciousness." *BMC Neuroscience* 5 (2004): article 42.
- Tononi, Giulio. *Phi: A Voyage from the Brain to the Soul*. Pantheon Books, 2012.
- Albantakis, Larissa, et al. "Integrated Information Theory (IIT) 4.0: Formulating the Properties of Phenomenal Existence in Physical Terms." *PLOS Computational Biology* 19, no. 10 (2023): e1011465.
- Doerig, Adrien, Aaron Schurger, Kathryn Hess, and Michael H. Herzog. "The Unfolding Argument: Why IIT and Other Causal Structure Theories Cannot Explain Consciousness." *Consciousness and Cognition* 72 (2019): 49–59.

### Related Concepts and Figures

- *The Hard Problem of Consciousness* — explaining the presence and quality of experience.
- *Panpsychism* — consciousness beyond familiar biological subjects.
- *Physicalism* — causal organization and competing identity claims.
- *Claude Shannon* — statistical information, distinct from IIT's intrinsic measures.
- *Anil Seth* — a different explanatory strategy in consciousness science.

### Related Writing on This Site

- Mind, Matter, and Meaning — the site author's philosophical argument, separate from this reference account.

### My View

I admire IIT's nerve and its refusal to reduce consciousness to report or behavior. IIT 4.0 treats first-person structure as a constraint on physical theory and develops a causal formalism far more exact than the old shorthand of a single Φ score. Its clinical descendants may also provide useful measures of conscious level, though those proxies do not directly measure or independently validate metaphysical Φ.

The book nevertheless parts company with IIT at a deeper point than the 2023 controversy. IIT analyzes intrinsic difference-making within a system whether or not that system presents a world. A causally irreducible lattice could therefore receive a rich conscious structure while representing nothing beyond itself. Integration may help identify a subject and explain poising, but it cannot stand alone as phenomenal character. On the house identity claim, phenomenal character consists in world-presenting content of the right kind; integration enters as part of the organization through which such content becomes one subject's presented world.

IIT's widespread-consciousness implications sharpen the cost. Expander-style systems and simple integrated structures may receive consciousness despite lacking recognizable world-directed cognition, while behaviorally sophisticated feed-forward systems receive none. The book can learn from IIT's account of causal integration without borrowing its criterion. Tononi locates consciousness in intrinsic causal structure; the book locates it in suitably organized world-presentation and treats the residual hard-problem gap as epistemic.

### Alan Turing

Alan Turing (1912–1954) developed a mathematical account of computation and proposed an influential approach to evaluating machine intelligence. His universal-machine construction concerns what can be computed; his imitation game concerns evidence for thinking. Neither claim should be treated as a theorem that computation alone produces consciousness.

### Major Works

"On Computable Numbers, with an Application to the Entscheidungsproblem" was published in 1936–1937. It introduces the abstract machines now bearing Turing's name and establishes a limit on general mechanical decision procedures. "Computing Machinery and Intelligence" (1950) proposes the imitation game, examines objections to machine thinking, and considers how machines might learn.

### Machines and Computability

A Turing machine has a tape, a finite control, and rules for reading, writing, changing state, and moving along the tape. Turing uses this idealization to analyze the operations available in an effective mechanical procedure. The unrestricted tape belongs to the mathematical model; a real computer has finite memory and runs subject to physical limitations.

A **universal machine** can simulate another machine when supplied with an encoding of its description and input. This gives a precise sense in which a single general-purpose device can execute many different procedures. It does not say that every procedure is computable or that every computation finishes.

Turing's results show that no general mechanical procedure can decide the full class of problems targeted by the Entscheidungsproblem. They also exhibit computations whose termination cannot be decided by one universal algorithm.

### The Church–Turing Thesis

The Church–Turing thesis identifies effective calculability with Turing computability and equivalent formal models. Mathematical equivalences among those models can be proved; the identification with the informal notion of effective procedure has a different status.

The thesis is not, as stated, a claim that all physical processes are computable, that every computation is efficient, or that all thought consists in computation. Physical and complexity-theoretic extensions require additional assumptions. Multiple physical implementations of a procedure likewise do not alone establish multiple realization of every mental property.

### The Imitation Game

Turing begins the 1950 paper with a game in which an interrogator communicates remotely with a man and a woman and tries to identify them. He then considers replacing one participant with a machine. The familiar later "Turing test" emphasizes a text-based comparison between a human and a machine; the relation between this formulation and the original gender game is an interpretive issue.

The point is to replace a disputed definition of thinking with an assessable task. Remote textual communication limits the role of appearance and physical resemblance. Turing predicts that machine performance will make attributing thought increasingly ordinary; he does not specify the absolute requirement that every possible judge must always be deceived.

### Consciousness and Other Objections

Turing considers the claim that a machine must feel what it does before being credited with thought. Demanding direct access to another being's experience would also threaten ordinary judgments about other people. His reply therefore treats behavior and interrogation as evidence while acknowledging the difficulty of consciousness. It does not establish that machines have no inner experience or that experience is irrelevant in every context.

Against the mathematical objection, he questions whether humans are themselves free from the limitations used to criticize machines. Against Lady Lovelace's objection that machines only do what they are instructed to do, he emphasizes the possibility of unexpected results and learning. He also discusses alleged disabilities, the continuity of nervous systems, and other objections. These discussions are arguments about proposed barriers, not proofs that any particular machine thinks.

### Learning Machines

Instead of programming an adult mind in all its detail, Turing proposes beginning with a child machine and educating it. He considers reward, punishment, changes in behavior, and processes analogous to selection. The proposal is broader than matching a fixed set of conversational answers: it concerns how capabilities could develop.

This aspect of the paper complicates readings that reduce Turing's entire philosophy of intelligence to deceptive conversation. His discussion includes organization, education, and the acquisition of capacities, even though the imitation game supplies his proposed public test.

### Criticisms and Continuing Significance

Critics argue that an input–output test could reward shallow imitation, misleading presentation, or a hypothetical lookup mechanism rather than understanding. Searle's Chinese Room questions the sufficiency of formal language performance. Other critics object that intelligent systems need not resemble humans in conversation.

Defenders treat demanding, flexible, extended performance as defeasible evidence rather than a metaphysical definition or infallible test. Different testing conditions matter: success in one constrained exchange is not success under every version of the game. The continuing debate

concerns what evidence establishes which capacities, not just whether a system can produce fluent text.

**Further Reading**

- Turing, A. M. "On Computable Numbers, with an Application to the Entscheidungsproblem." *Proceedings of the London Mathematical Society*, second series, 42 (1936–1937): 230–265.
- Turing, A. M. "Computing Machinery and Intelligence." *Mind* 59, no. 236 (1950): 433–460.
- Searle, John R. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3, no. 3 (1980): 417–424, followed by commentary and response.

**Related Concepts and Figures**

- *Computation and the Church-Turing Thesis* — computability and its extensions.
- *The Turing Test* — testing conditions and evidential claims.
- *Functionalism* — causal organization and mental states.
- *The Chinese Room* — a challenge to program sufficiency.
- *John von Neumann* — stored-program computing and the comparison with brains.

**Related Writing on This Site**

- Mind, Matter, and Meaning — the site author's philosophical argument, separate from this reference account.

**My View**

Turing is indispensable, and the book engages him at two distinct seams, taking care not to blame him for conflations made in his name.

First, the **Church–Turing thesis is not computationalism.** Turing told us which functions a machine can compute. He did not thereby tell us that minds *are* such computations, or that running the right computation suffices for a mind. Treating his thesis as if it answered the metaphysical question is one of the cleaner category errors in the field. The book separates the two explicitly (see *Computation and the Church-Turing Thesis* and *Simulation and Realization*). Turing the logician proved something true and precise; the overreach belongs to his readers.

Second, the **Turing test measures genuine behavioral competence without constituting understanding or consciousness by score alone.** Fluent performance remains defeasible evidence: stronger, more flexible, correctable behavior can raise the probability that a system integrates and uses content. That evidence must be read together with provenance, architecture, persistence, world engagement, and the boundary of the candidate system. Turing deserves credit for operationalizing a difficult question and for keeping consciousness separate. The overreach begins when successors treat indistinguishable text as proof of an inner life. In the LLM era, the imitation game is therefore neither a bar the book simply rejects nor a verdict it accepts; it is one source of evidence inside the wider assessment developed in Chapters 19, 24, and 25.

**Michael Tye**

Michael Tye's work examines phenomenal consciousness, perceptual representation, animal experience, and personal identity. His position developed through several distinguishable phases:

the PANIC account of the 1990s and 2000s, a revised materialist response to phenomenal knowledge, and a later panpsychist account. The later work retains a representational explanation of experiential differences while revising what explains the existence of phenomenality.

**Major Works and Phases**

*Ten Problems of Consciousness* (1995) develops the PANIC framework. *Consciousness, Color, and Content* (2000) refines it and examines transparency, inverted qualia, and other objections. *Consciousness and Persons* (2003) considers unity and personal identity. *Consciousness Revisited: Materialism without Phenomenal Concepts* (2009) rejects a strategy he had previously defended and emphasizes acquaintance. *Tense Bees and Shell-Shocked Crabs* (2017) investigates animal consciousness. *Vagueness and the Evolution of Consciousness* (2021), followed by "How I Learned to Stop Worrying and Love Panpsychism" (2024), develops the later view.

**PANIC and Phenomenal Character**

In the earlier framework, a state's phenomenal character is identified with **Poised, Abstract, Nonconceptual, Intentional Content**:

- **Poised:** the state stands ready to influence appropriate beliefs or desires directly. Actual attention or use on every occasion is unnecessary.
- **Abstract:** the relevant phenomenal content involves general properties rather than requiring particular objects as constituents. Matching appearances of distinct red objects need not differ simply because the objects differ.
- **Nonconceptual:** representing a property does not require possessing a concept for it. Experience can discriminate shades or shapes more finely than the subject can conceptualize or remember them.
- **Intentional:** the state represents conditions and has accuracy conditions; its content is not merely a pattern of internal activity.

On this account, early sensory processing may carry content without being appropriately poised for experience. Blindsight motivates the need to distinguish discrimination from phenomenal vision. These are components of a historical theory, not an unchanged specification of everything Tye subsequently endorsed.

**Transparency and Externalism**

When a person attends to a visual experience of a surface, Tye argues, they attend to the surface's presented color, shape, and other properties rather than to an additional colored mental object. He uses this transparency to motivate representationalism.

Critics argue that failing to notice an intrinsic experiential property does not prove its absence, and that inverted-spectrum cases might preserve content while changing experience. Tye challenges whether the proposed cases really hold relevant content fixed and offers externalist accounts of its determination. The inference from transparency to a complete identity theory remains contested.

"Homunculi Heads and Silicon Chips: The Importance of History to Phenomenology" argues that intrinsic similarity at one time need not guarantee phenomenal similarity. The history

through which organization acquires representational functions matters. The essay also considers functional replacement with silicon chips, distinguishing an acquired history from an arbitrary present-day organization.

### Knowledge and Acquaintance

Tye's responses to Mary's Room develop over time. His earlier work discusses abilities, perceptual representation, and ways of thinking about qualities. *Consciousness Revisited* rejects the special phenomenal-concept strategy and draws on the distinction between acquaintance with something and descriptive knowledge about it.

On this materialist approach, encountering red can provide a kind of epistemic access absent from Mary's exhaustive scientific education without revealing an additional nonphysical property. Critics ask whether acquaintance can explain the apparent discovery or whether it already presupposes the relation requiring explanation. It is misleading to reduce all of Tye's responses to one unchanging "new ability, no knowledge" answer.

### The Later Panpsychist Account

Tye argues that phenomenal consciousness is essentially sharp: being conscious cannot itself be genuinely borderline, although our evidence can leave us uncertain. Gradual evolution and broadly reductive accounts then appear unable to explain a nonarbitrary arrival of consciousness. His conclusion introduces a fundamental phenomenal determinable called **consciousness***.

The determinable is not simply one tiny, fully specified humanlike experience. Representational content explains determinate experiential character, while consciousness* explains the occurrence of phenomenality. In the 2024 account, he proposes transfer of consciousness* from microstates to complex states when an appropriate global-workspace condition is definitely satisfied. He does not conclude that every table, rock, or arbitrary aggregate is a conscious subject.

This revises the earlier reductive ambition while retaining a representational account of differences between experiences. Tye explicitly distinguishes his newer account from PANIC.

### Objections and Replies

Critics distinguish the apparent sharpness of our concept from sharpness in nature, and question whether gradual development necessarily makes the property vague. Tye replies that his argument concerns the nature disclosed by phenomenal consciousness, not merely a convenient feature of our vocabulary.

A further objection asks why microlevel consciousness* transfers to a whole under the proposed condition and how that establishes one subject. Tye presents the global-workspace proposal as a solution to combination; critics regard the transfer principle as needing independent explanation. His view is therefore not adequately assessed by either repeating the standard microsubject-summing objection unchanged or treating the proposed solution as settled.

### Further Reading


- Tye, Michael. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind.* MIT Press, 1995.

- Tye, Michael. *Consciousness, Color, and Content*. MIT Press, 2000.
- Tye, Michael. *Consciousness Revisited: Materialism without Phenomenal Concepts*. MIT Press, 2009.
- Tye, Michael. "Homunculi Heads and Silicon Chips: The Importance of History to Phenomenology." In *Blockheads! Essays on Ned Block's Philosophy of Mind and Consciousness*, edited by Adam Pautz and Daniel Stoljar, 545–570. MIT Press, 2019.
- Tye, Michael. *Vagueness and the Evolution of Consciousness: Through the Looking Glass*. Oxford University Press, 2021.
- Tye, Michael. "How I Learned to Stop Worrying and Love Panpsychism." *Journal of Consciousness Studies* 31, nos. 9–10 (2024): 10–28.

**Related Concepts and Figures**

- *Intentionalism* — representational accounts of phenomenal character.
- *Transparency* — attention to experience and its presented objects.
- *Mary's Room* — experiential and descriptive knowledge.
- *Panpsychism* — fundamental phenomenality and combination.
- *Ned Block* — objections involving functional organization and qualia.

**Related Writing on This Site**

- Mind, Matter, and Meaning — the site author's philosophical argument, separate from this reference account.

**My View**

Michael Tye is a major contemporary interlocutor, not the authority for the book's position. His representationalist work develops one of the clearest modern versions of the identity claim and its externalist consequences. Dretske, Harman, Moore, Crane, Block, Millikan, Byrne and Hilbert, and others supply different parts of the argument and its constraints. The resulting synthesis belongs to this book.

I take the full identity claim: a conscious state's phenomenal character consists in its complete world-presenting intentional profile. That commitment converges with Tye's earlier representationalism while going beyond PANIC in how the book separates content fixing, phenomenal character, the conscious threshold, and subject individuation. PANIC remains one historically important hypothesis about perceptual form and poising, not a complete theory or an item accepted on Tye's authority.

On Tye's late panpsychism: I do not follow him across. I reject the inference from our inability to frame borderline consciousness to a sharp phenomenal joint in nature, while recognizing that Tye treats sharpness as an a-priori disclosure of the property rather than merely a feature of our concept. His consciousness* has a real theoretical job — grounding that phenomenality occurs — but the view pays for a fundamental phenomenal determinable plus a transfer principle from microstates to an organized whole. The book pays instead for a strong a-posteriori identity with no conceptual derivation. Its preference remains comparative and conditional, not a claim that only Tye's view contains a primitive. See *Panpsychism*.

The disagreements reach beyond panpsychism. The book treats phenomenal mode as introspectively accessible, keeps the conscious-threshold mechanism unsettled, gives subject individuation its own integration-and-intervention test, and uses an acquaintance-recognition account without adopting Tye's mature materialism wholesale. These are substantive differences inside a shared representationalist territory. See *Intentionalism*, *Transparency*, and *Embodiment*.

**John von Neumann**

John von Neumann (1903–1957) worked on mathematics, quantum theory, economics, computing, and automata. His report on the EDVAC helped disseminate stored-program computer design. His unfinished final book, *The Computer and the Brain*, compares engineered computation with the nervous system while emphasizing differences in their organization. That comparison is neither a proof that brains are conventional digital processors nor a proof that artificial minds are impossible.

**Major Works**

- *Mathematical Foundations of Quantum Mechanics* (1932) — the mathematical framework of quantum theory.
- *Theory of Games and Economic Behavior* (with Oskar Morgenstern, 1944) — strategic interaction and economic behavior.
- "First Draft of a Report on the EDVAC" (1945) — logical organization of a stored-program computer.
- *The Computer and the Brain* (1958, posthumous) — physical and organizational comparisons between computers and nervous systems.
- *Theory of Self-Reproducing Automata* (1966, edited and completed by Arthur W. Burks) — formal organization capable of construction and reproduction.

**Stored Programs and Historical Attribution**

In a stored-program design, encoded instructions are held in memory rather than supplied only through rewiring or a separate fixed control arrangement. The machine can read instructions as part of its operation; instructions and data can share a representational medium. A general-purpose design can then execute different procedures without rebuilding the hardware for each one.

The EDVAC report describes arithmetic, control, memory, input, and output as organized components. It abstracts from particular engineering devices while also examining physical implementation. The report's influence helps explain the expression **von Neumann architecture**.

The design emerged from collaborative work, including J. Presper Eckert, John Mauchly, Arthur Burks, Herman Goldstine, and von Neumann. Sole authorship of the circulated report should not be confused with sole invention of the stored-program concept. Historical priority is more complex than the architectural name suggests.

Nor does a real stored-program computer literally run every possible computation without limits. Memory, time, available instructions, and physical reliability constrain what any particular device

can do. Turing's universal machine provides a mathematical comparison, not a guarantee of unlimited hardware.

### Self-Reproducing Automata

Von Neumann investigates how a machine could construct another machine, including one of its own kind. The organization must use a description in construction and copy that description for the offspring. This separates interpreting a specification from copying it, giving a formal account of how reproduction can preserve an organization across generations.

The project concerns constructive organization rather than proof that a machine could spontaneously create a living organism from nothing. Its detailed theory was published posthumously through Burks's editorial work. Connections with biological reproduction are instructive analogies, not a complete theory of life or consciousness.

### The Computer and the Brain

Von Neumann compares the speed, size, energy use, reliability, memory, and signaling of engineered machines with nervous systems. A neuron's all-or-none firing suggests a digital element, but its operation within a biological system also involves continuous variables and chemical processes. The comparison therefore does not support a simple all-digital or all-analog classification.

A key contrast concerns precision. Conventional numerical machines rely on highly exact operations, whereas neural signaling works with comparatively low precision and statistical patterns. Slower components and distributed organization can still support complex abilities. Von Neumann asks how the nervous system compensates for limitations that would be serious in an ordinary serial calculation.

His concluding discussion suggests that the nervous system's operative "language" differs from familiar formal logic and arithmetic: it has less logical and arithmetical depth and a strongly statistical character. This is a conjecture about organization grounded in the comparison, not a discovery of the brain's complete computational scheme.

### Implications and Limits

*The Computer and the Brain* argues for investigating what a nervous system actually does instead of treating one engineered architecture as a universal template. It does not reject mathematical modeling of neural activity or the possibility that other artificial architectures could reproduce relevant capacities.

A computationalist can accept the architectural differences while arguing that the right abstract organization is what matters. A biological or dynamical theorist can argue that some physical features omitted by conventional models are indispensable. Von Neumann's comparison supplies constraints and questions for both sides; it does not resolve the dispute by itself.

### Criticisms and Later Interpretation

The empirical estimates reflect mid-twentieth-century knowledge and available machines. Modern parallel hardware and neural models make "brain versus computer" a less uniform contrast

than the book's historical comparison can suggest. Its unfinished form also limits how confidently a completed position can be reconstructed.

Conversely, later technical developments do not automatically answer the explanatory questions about precision, memory, and organization. The enduring problem is which properties of an implementation explain a capacity, and which can vary without changing it. That question extends beyond a contrast between carbon and silicon.

**Further Reading**

- von Neumann, John. "First Draft of a Report on the EDVAC." Moore School of Electrical Engineering, University of Pennsylvania, June 30, 1945.
- von Neumann, John. *The Computer and the Brain*. Yale University Press, 1958.
- von Neumann, John. *Theory of Self-Reproducing Automata*. Edited and completed by Arthur W. Burks. University of Illinois Press, 1966.

**Related Concepts and Figures**

- *Computation and the Church-Turing Thesis* — abstract computation and physical implementations.
- *Functionalism* — organization and multiple realization.
- *Simulation and Realization* — which properties an implementation reproduces.
- *Carbon Chauvinism* — arguments about material restrictions on minds.
- *Alan Turing* — universal computation and machine intelligence.

**Related Writing on This Site**

- Mind, Matter, and Meaning — the site author's philosophical argument, separate from this reference account.

**My View**

Von Neumann is the right figure to hold the line the book draws between *what computation can be realized on many substrates* and *what a mind can be realized on*. The first is true, and his architecture is the monument to it. The second does not follow, and von Neumann himself declined to draw it. Read *The Computer and the Brain* and you find not a triumphalist reduction of mind to machine but a meticulous catalog of the ways the brain refuses the comparison: its parallelism, its analog gradations, its tolerance for noise, its shallow logical depth.

I value this because the strongest form of the computational picture likes to claim von Neumann as a founding father. The claim is half true at best. He gave us the universal machine in practice; he did not give us the universal *mind*. But the book's target is not digitality itself. An abstract program omits the world-involving history and organization relevant to mentality, while a concrete digital or virtual system could in principle realize content-bearing producer-consumer mechanisms, persistent integration, and poised perceptual organization. Von Neumann's caution supplies the right lesson: computational architecture does not settle what mental capacities an implementation realizes. Cross-material recurrence remains possible, but it must be shown capacity by capacity rather than inferred from program identity.

### Ludwig Wittgenstein

Ludwig Wittgenstein (1889–1951) investigated logic, language, knowledge, and psychological concepts. The early *Tractatus Logico-Philosophicus* and the later *Philosophical Investigations* differ substantially, but their relation is also a subject of interpretation. His later work examines language in use, rule-following, and the temptation to explain meaning through private mental objects. It does not offer one conventional explanatory theory of mind.

### Major Works and Development

The *Tractatus* appeared in German in 1921 and in a bilingual book edition in 1922. *The Blue and Brown Books* contains dictations from the 1930s, published in 1958. *Philosophical Investigations* appeared posthumously in 1953. *Zettel* (1967) collects remarks; *On Certainty* (1969) investigates the background of knowledge and doubt. These works differ in composition and editorial history, rather than forming a series of completed treatises published by their author.

### The Early Picture Account

In the *Tractatus*, propositions picture possible states of affairs by sharing a logical form with what they represent. Their elements are correlated with objects, and their arrangement represents a possible arrangement in the world. Truth and falsity concern whether the represented possibility obtains.

The work also distinguishes what can be said from what is shown by meaningful representation. Its remarks on logic, ethics, and the limits of language resist a straightforward summary as a theory that every meaningful sentence names an observable object. How literally to treat its apparent metaphysical commitments remains disputed among interpreters.

### Meaning, Use, and Language-Games

The *Investigations* criticizes the assumption that words work uniformly as names and that understanding must consist in an inner accompaniment. Language includes asking, ordering, describing, joking, calculating, and many other activities. Wittgenstein calls these organized activities **language-games**, embedded in forms of life.

Section 43 qualifies the formulation "meaning is use": it applies to a large class of cases, not to every case without exception. Describing use also means describing standards of correct application, circumstances, responses, and training. It is not merely counting word frequencies or substituting observable behavior for every question about meaning.

**Family resemblance** challenges the demand that every concept pick out one feature common to all its instances. Games, for example, can be related through overlapping similarities rather than a single essence. The proposal concerns how a concept operates; it does not entail that no concepts have precise definitions.

### Following a Rule

The rule-following discussion asks how a rule guides a new application. An interpretation can itself be interpreted, so adding interpretations cannot indefinitely supply what makes an application correct. Wittgenstein points to training, established practice, and ways of going on that are not always mediated by an explicit further interpretation.

The significance of communal agreement is debated. Some readings emphasize public practices; others resist making community approval the ultimate truth-maker for correctness. Saul Kripke's influential skeptical reconstruction is one interpretation, not an uncontested statement of Wittgenstein's doctrine. Rule-following need not be reduced either to an inner formula or to a majority vote.

**Private Language and Sensation**

A private language in the targeted sense is not an ordinary diary, a secret code, or speech used while alone. Its signs would refer to sensations whose identification and correctness conditions were available only to their speaker.

In section 258, a person tries to fix the meaning of "S" by inwardly attending to a sensation and later consulting memory. The difficulty is that seeming to remember the association correctly supplies no independent distinction between correctness and the impression of correctness. This concerns standards for the proposed sign, not the conclusion that sensations do not exist.

The "beetle in a box" example in section 293 makes a related point. If everyone has an inaccessible box and calls its contents "a beetle," what those inaccessible contents are cannot play the explanatory role assigned to them in the shared word's use. Ordinary sensation language has learned roles tied to expression, circumstances, and response; it is not simply a collection of privately attached labels.

**Psychological Concepts and Method**

Wittgenstein rejects both the picture of a hidden object behind every mental expression and the behaviorist denial of inner life. Pain behavior, for example, can enter the grammar of pain language without pain being identical to a movement or utterance. First-person avowals and third-person judgments need not have the same role or evidential basis.

His therapeutic method seeks to clarify how language generates philosophical problems. Critics ask whether clarification can settle substantive disputes about mind, and whether public criteria adequately respect subjective experience. Defenders reply that exposing a misleading question can be philosophical progress without eliminating the phenomenon that prompted it.

Later externalists and philosophers of mind draw on these investigations, but Wittgenstein should not be credited with endorsing their subsequent causal, biological, or computational theories.

**Further Reading**


- Wittgenstein, Ludwig. *Tractatus Logico-Philosophicus*. Translated by C. K. Ogden. Kegan Paul, Trench, Trubner & Co., 1922. German publication, 1921.
- Wittgenstein, Ludwig. *Philosophical Investigations*. Translated by G. E. M. Anscombe. Basil Blackwell, 1953.
- Wittgenstein, Ludwig. *The Blue and Brown Books*. Basil Blackwell, 1958.
- Wittgenstein, Ludwig. *On Certainty*. Edited by G. E. M. Anscombe and G. H. von Wright; translated by Denis Paul and G. E. M. Anscombe. Basil Blackwell, 1969.

**Related Concepts and Figures**

- *Semantic Externalism* — later theories of meaning beyond the individual.
- *Inner Speech* — public language and silent verbal activity.
- *Anti-Reification* — the explanatory role of mental objects.
- *Wilfrid Sellars* — a distinct account of inner-episode concepts.
- *Hilary Putnam* — later externalism and natural realism.

**Related Writing on This Site**

- Mind, Matter, and Meaning — the site author's philosophical argument, separate from this reference account.

**My View**

Wittgenstein supplies two visible moves that carry real argumentative work in the current book without becoming its house authority. In Chapter 17, *The Assembled Self*, the private-language argument shows that the vocabulary of inwardness answers to public criteria: privileged access to experience survives, but no private act of inner ostension fixes a language's meanings. Sellars and Vygotsky then develop different explanatory routes from public speech to inner-episode talk.

Chapter 19, *The Problem of Other Minds*, draws on Wittgenstein's claim that we see emotion. Concern, grief, and attention can be perceived directly in expressive, world-directed activity rather than inferred as hidden objects behind behavior. The perception remains defeasible: acting, concealment, context, and correction can defeat it. Direct social perception therefore rejects neither fallibility nor the reality of another person's first-person life, and it is not behaviorism.

The private-language argument also supplies a precursor to semantic externalism. Wittgenstein blocks the picture in which meaning is fixed wholly inside one skull and locates meaning in public use and practice. Kripke, Putnam, and Burge then add the causal-historical, environmental, and social considerations through which those practices reach beyond their signs. Chapter 16 presents this as a cumulative development, not as a staged break with Wittgenstein and not as a claim that he endorsed the later externalist theory. His meaning-as-use framework enters the AI question only in the book's qualified sense: meaningful use is world-reaching, reference-involving use within practices of correction, not pure inferential role and not an alternative to reference.

The book takes Wittgenstein's account of public use as a starting point and places the later externalist considerations beside it. The constructive account that follows belongs to the book's synthesis; Wittgenstein need not receive either credit or blame for claims he did not make.

**Arnold Zuboff**

Arnold Zuboff is a philosopher, long associated with University College London, who works on personal identity and the mind–body problem. He is best known for two lines of argument that he keeps separate. One is **universalism**, the claim that the boundary between one conscious being and another is an illusion of perspective, so that there is a single subject of all experience. The other is the **replacement argument**, which defends functionalism by asking what could happen to experience when part of a brain is swapped for a causally equivalent gadget. He also reports that his unpublished work on self-locating examples was the origin of what became known as the

Sleeping Beauty problem; Adam Elga's 2000 paper credits Robert Stalnaker with first learning of such examples from it.

**Major Works**

- "Moment Universals and Personal Identity," *Proceedings of the Aristotelian Society* 78 (1977–78) — the earlier identity paper, which argues that each moment of experience is a universal and that we are not persons in the ordinary sense.
- "The Story of a Brain," in Douglas Hofstadter and Daniel Dennett (eds.), *The Mind's I* (1981) — the spread-brain thought experiment.
- "One Self: The Logic of Experience," *Inquiry* 33 (1990) — the settled statement of universalism.
- "What Is a Mind?," *Midwest Studies in Philosophy* 19 (1994) — the worked version of the replacement argument.
- "Thoughts about a solution to the mind-body problem," *Think* 6 (2008) — a short restatement of the replacement argument, built around the Necker cube.
- *Finding Myself: Beyond the False Boundaries of Personal Identity* (Philosophy Documentation Center, 2025), with a foreword by Thomas Nagel — the book-length development of universalism, published open access.

**Universalism**

The usual view holds that you are one particular conscious being, picked out by a body, an origin, or a continuous history. Zuboff argues that this view smuggles in an assumption nothing supports. Before you consider any objective description of yourself, you already know that you are the being whose experience is *immediate*, the one that seems to sit at the center. Every conscious being has experience with exactly that first-person immediacy. On Zuboff's account, each experience seems to its owner to be the only one that is mine because it is cut off from the contents of the others, and nothing more is needed to explain that feeling.

His arguments come in several kinds.

- **Brain bisection.** Imagine anaesthetizing the corpus callosum so that the two hemispheres briefly work apart. Each side would have immediate experience, and each would feel like the whole of it. Zuboff concludes that the immediacy cannot belong more to one side than to the other, so that being "mine" extends to experiences that cannot reach each other. He credits the structure of the case to a variation on Derek Parfit's 1971 paper "Personal Identity."
- **The brain-exchange case.** Two exactly duplicate brains are induced to have exactly the same pattern of experience, and parts are swapped between them one piece at a time. Zuboff argues that no stage of the exchange can mark a switch of places, and that the end result is indistinguishable from simply picking the brains up and putting them down in each other's place. He reads this as showing that qualitatively identical experiences in two brains are one experience, not two.
- **The probability argument.** If your existence required that the one right sperm cell win the race in the begetting of you and of each of your ancestors, then your being here would be fantastically improbable. Zuboff argues that the only view that does not make your existence an incredible coincidence is universalism, on which you would have existed whichever organisms

happened to arise. He presents this as parallel to inferring that an awakening came from an easy game rather than from a game that required a thousand coin tosses to match a secret code.

A consequence Zuboff draws is that self-interest should extend to all conscious beings, and that the death of one organism does not end the subject, since there is only one.

**Development**

Zuboff's view changed in stages, and the stages should not be run together.

- **1978.** In "Moment Universals and Personal Identity" he argues for *unification*, the claim that exactly repeated experiences make up one experience, and for the conclusion that each moment of experience is a universal. He concludes there that there are no proper continuing persons, and he treats the sense of being a person as a product of evolution.
- **1983.** In *Finding Myself* he writes that he discovered universalism in 1983. It is the further step that all experience is equally immediate and equally mine, and it replaces the earlier stopping point.
- **1990.** "One Self" develops universalism and its supporting arguments.
- **2025.** *Finding Myself* sets out the whole case, including the probability argument and a treatment of the Sleeping Beauty problem. Zuboff notes that Elga's 2000 paper credits Robert Stalnaker with having first learned of such examples from Zuboff's unpublished work, which appeared in 1990 as "One Self."

**The Replacement Argument**

The second line of argument concerns what fixes the character of an experience. Zuboff calls the opposing view **intrinsicism**: the idea that some intrinsic property of a bit of neural processing, perhaps chemical or biological, or perhaps a distinct mental property, is what gives an experience its character. Functionalism says that only the causal role matters.

The argument runs as a story. Part of your brain, say the left visual cortex, is replaced by a gadget that keeps exactly the same causal relations to the rest of the brain, though its intrinsic properties are radically different. The intrinsicist must tell the story in one of two ways.

- In **"A Change Expressed,"** your vision on one side really is altered and you complain about it. This cannot be right, because the gadget is stipulated to be a perfect functional equivalent, so nothing you say, do, or think could differ.
- In **"A Change Unexpressed,"** your vision on that side really is altered, but you speak and behave exactly as if it were normal. Zuboff argues that this is not coherent either, because a changed experience could not leave the speech, thought, and behavior grounded in that experience unchanged.

He concludes that functioning preserves the phenomenal character of experience, and that the argument establishes functionalism a priori. "The Story of a Brain" (1981) is an earlier piece in the same neighborhood. It follows a brain divided into hemispheres, then neurons, and asks what would remain of the physical conditions on which a neural theory of experience relies. The story ends with the theory seeming to undermine itself, since it appears to provide experience so cheaply that no experiencer could trust the physiology on which the theory rests. Zuboff says in

1994 that his own position blends Gilbert Harman's functionalism with Thomas Nagel's insistence on fidelity to the subjective character of experience.

The two lines of argument are logically separate. Preserving experience under replacement does not show that every conscious being is the same subject, and universalism does not by itself fix what makes a particular experience the one it is.

**Reception and Criticism**

Nagel's foreword to *Finding Myself* grants that "both the conclusion and the argument will certainly seem incredible, even outrageous, to many readers." He nonetheless calls the book "a philosophical contribution of the first order" and says that the relation of mind to world is so mysterious that no view, however radical, should be dismissed out of hand.

Critics press several points. The probability argument depends on contested principles for reasoning about one's own identity and about self-locating evidence. The bisection and exchange cases rest on thought-experiment verdicts that a reader may resist. Nick Bostrom argues that duplicated experience adds further experience instead of merging with the original, a dispute Zuboff addresses directly in *Finding Myself*. On the replacement argument, a defender of intrinsic or biological views can deny that the stipulated functional equivalence leaves experience unchanged, or can reject the step from unchanged reports to unchanged experience. David Chalmers's "fading qualia" and "dancing qualia" arguments are close relatives of the replacement argument, though Chalmers draws a different overall picture of the mind from them.

**Related Concepts and Figures**

- *The Self* — what "the self" can name, and how the personal-identity question differs from the subject question.
- *Functionalism* — the view that mental states are individuated by causal role.
- *Qualia* — phenomenal character and what it might depend on.
- *Physicalism* — the wider family of views that Zuboff's functionalism belongs to.
- *Thomas Nagel* — graduate supervisor, author of the foreword to *Finding Myself*, and the author of the "I am Thomas Nagel" passage on token-reflexive facts that Zuboff discusses in "One Self."
- *Gilbert Harman* — graduate supervisor, and the functionalist influence Zuboff names.

**Further Reading**


- Zuboff, Arnold. "Moment Universals and Personal Identity." *Proceedings of the Aristotelian Society* 78 (1977–78): 141–155.
- Zuboff, Arnold. "The Story of a Brain." In *The Mind's I: Fantasies and Reflections on Self and Soul*, edited by Douglas R. Hofstadter and Daniel C. Dennett, 202–212. Basic Books, 1981.
- Zuboff, Arnold. "One Self: The Logic of Experience." *Inquiry* 33, no. 1 (1990): 39–68.
- Zuboff, Arnold. "What Is a Mind?" *Midwest Studies in Philosophy* 19 (1994): 183–205.
- Zuboff, Arnold. "Thoughts about a Solution to the Mind-Body Problem." *Think* 6, no. 17 (2008): 159–171.
- Zuboff, Arnold. *Finding Myself: Beyond the False Boundaries of Personal Identity*. Foreword by Thomas Nagel. Philosophy Documentation Center, 2025. Open access (CC BY-NC-ND): https://www.pdcnet.org/wp/wp-content/uploads/2025/12/Zuboff_MSP-Supplement_2025.pdf.

- Parfit, Derek. "Personal Identity." *The Philosophical Review* 80, no. 1 (1971): 3–27.
- Elga, Adam. "Self-Locating Belief and the Sleeping Beauty Problem." *Analysis* 60, no. 2 (2000): 143–147.

**Related Writing on This Site**

- Mind, Matter, and Meaning — the site author's philosophical argument, separate from this reference account.

# The Concepts — A Reader's Glossary

*A reader's glossary of the book's working concepts — each entry a short standalone treatment, written to be read in any order.*

### Anti-Reification

Anti-reification is a methodological warning against treating an abstraction, activity, relation, or explanatory model as a separate thing merely because language gives it a noun. Saying that traffic has slowed does not introduce an additional object alongside the vehicles. Saying that someone possesses intelligence need not locate a substance called intelligence inside that person.

The warning has no single agreed formulation in philosophy of mind. It can support ordinary conceptual clarification, a theory of mental processes, or a more revisionary criticism of mental vocabulary. Those projects should be distinguished: identifying a misleading picture does not by itself establish which positive theory should replace it.

### Language and Ontological Commitment

A noun can name an object, a property, an event, a capacity, or a relation. The grammatical form alone does not decide which. Gilbert Ryle's *The Concept of Mind* (1949) develops this concern through the idea of a **category mistake**: assigning something to a logical category to which it does not belong. His visitor who sees a university's colleges and libraries and then asks to see the university treats the organization as another building. Ryle applies a related diagnosis to a conception of mind as a second, hidden mechanism alongside the body.

Ludwig Wittgenstein's *Philosophical Investigations* (1953) examines how misleading analogies between uses of words can generate philosophical puzzles. His treatment of sensation language questions whether naming a sensation works just like pointing to a private object. Neither approach amounts to the general rule that whatever lacks a public location must be unreal.

The distinction between a **representational vehicle** and its **content** supplies another example. A map consists of paper and ink, but its depiction of a winding road does not make the paper winding. Applying a property of what a representation concerns to the representing process can produce a mistaken account of that process.

### Applications in Philosophy of Mind

Critics of an inner-theater model argue that explaining vision by introducing an internal picture and a spectator leaves perception unexplained: the spectator now needs its own account of seeing. Daniel Dennett's *Consciousness Explained* (1991) attacks the idea of a privileged brain location where all information arrives for presentation to a central observer. This criticism concerns a particular explanatory architecture; it does not prohibit neural representations or deny that people have experiences.

Intentionalists make a related argument about illusion and imagery. An experience as of a bent stick can represent bentness without containing a bent mental object. A sense-datum theorist may resist this account and defend an immediate object of awareness. Calling that object a reification does not yet answer the theorist's argument from perceptual appearance. The disagreement concerns what an experience involves, not whether the word *experience* may be used.

Other applications concern information and computation. Information can designate a statistical relation, a meaningful message, or a physical encoding. Moving between these senses without an argument risks treating a measure as a transferable substance or semantic content as an automatic consequence of correlation. Yet a careful theory may identify genuine physical mechanisms that encode and process information. Anti-reification need not entail that all computational descriptions are merely arbitrary interpretations.

### Objections and Limits

The principal objection is that the charge can assume the ontology it purports to establish. A property realist can acknowledge that pain is not an object on an internal stage while maintaining that experiences possess real qualitative properties. A dualist can reject a literal inner spectator without abandoning dualism. A process theorist, in turn, must explain which processes occur and how they account for the phenomena.

Defenders of the method therefore distinguish **diagnosis** from **elimination**. Replacing talk of an inner item with a specification of activity may preserve the phenomenon. It may also leave a substantive question unanswered. The test concerns explanatory gain: does the reformulation clarify the relevant distinctions and evidence, or simply rename the difficulty?

Anti-reification is most useful when it identifies a particular invalid inference between grammar, phenomenology, and ontology. It supplies no general proof of physicalism, intentionalism, or the unreality of abstract entities.

### Further Reading

- Ryle, Gilbert. 1949. *The Concept of Mind*.
- Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Translated by G. E. M. Anscombe. Blackwell.
- Dennett, Daniel C. 1991. *Consciousness Explained*. Little, Brown.

### Related Writing on This Site

- *Mind, Matter, and Meaning* — Gordon Swobe's book presents the author's argument about mind, consciousness, and meaning.

### My View

Anti-reification is the methodological foundation of everything I argue. Whenever a philosophical term (mind, qualia, consciousness, the self, meaning) generates intractable puzzles, my first diagnostic move is to ask what pattern, process, or relation the term abbreviates, rather than what inner object it names. Problems that survive this translation are real; problems that dissolve were artifacts of the picture.

### Behaviorism

Behaviorism covers several approaches that give behavior a central place in the study or analysis of mind. Some concern the methods of psychology; others concern what mental concepts mean or what mental states consist in. A demand for observable evidence does not by itself imply that private experiences do not occur.

Behaviorist psychology became influential in the early twentieth century partly in response to disagreements over introspection. Its researchers sought publicly testable accounts of learning and conduct. Philosophical behaviorists pursued a related but distinct question: whether mental vocabulary requires reference to a private realm hidden behind action.

### Main Varieties

**Methodological behaviorism** restricts psychological inquiry to publicly observable behavior and its conditions, or treats such evidence as the basis for scientific claims. It can remain agnostic about unobservable mental events. John B. Watson's program helped make prediction and control of behavior central scientific aims.

**Analytical or logical behaviorism** proposes that mental statements can be analyzed in terms of behavior, behavioral dispositions, and relevant circumstances. A disposition extends beyond what someone currently does: fragile glass need not be breaking, and a person disposed to complain of pain may currently remain silent.

**Ryle's philosophical approach** challenges the Cartesian division between an external body and an internal, immaterial mechanism. In *The Concept of Mind* (1949), Gilbert Ryle describes many mental capacities through patterns of intelligent conduct. His position should not be reduced to a claim that every mental term has a simple stimulus-response translation; its treatment of skills, motives, and dispositions is more complex, and its classification as behaviorism remains interpretively disputed.

**Skinner's radical behaviorism** also differs from a blanket exclusion of the private. It allows private events, including sensations and covert behavior, while resisting their use as independent inner causes that terminate explanation. B. F. Skinner emphasizes learning histories, reinforcement, and environmental contingencies. Radical behaviorism and methodological behaviorism therefore disagree about how private events enter a science of behavior.

### Reasons for the Approach

Behavioral evidence makes psychological hypotheses open to shared investigation. It also helps explain how people learn and apply mental concepts: expressions of pain and displays of understanding provide recognizable criteria for their use. Behaviorist learning theories offer experimentally tractable accounts of how consequences alter behavior.

For a philosophical behaviorist, these connections may be constitutive rather than merely evidential. Describing someone as knowing how to ride a bicycle involves a capacity to perform under appropriate circumstances, not simply a hidden event whose relation to riding is accidental. Critics can accept this insight about skills while denying that it extends to every kind of sensation or thought.

### Objections and Replies

**Pain without expression.** A stoic may suffer without showing it. An actor may display pain without suffering. Defenders reply that a brief performance does not reproduce a complete set of dispositions: the actor and sufferer behave differently across a sufficiently wide range of circum-

stances. The critic's stronger case stipulates a sufferer whose pain-related dispositions have been suppressed. Whether such a case is coherent tests the analysis itself.

**Interdependence of mental states.** Whether a belief produces an action depends on other beliefs and desires. Believing that rain is coming leads someone to carry an umbrella only if they want to stay dry and believe the umbrella will help. An analysis of belief that invokes those further mental states has not yet translated mental vocabulary into purely behavioral terms. Functionalists respond by specifying a network of internal causal roles rather than analyzing each state separately.

**Qualitative character.** Critics argue that an account of dispositions can omit how pain or color experience feels. Behaviorists question whether this alleged remainder has been characterized independently of the practices through which experience is identified and described. The dispute goes beyond the availability of behavioral evidence to concern what that evidence can establish.

### Continuing Influence

Cognitive and neuroscientific approaches displaced behaviorism as the dominant framework in much psychology, but behavioral research and intervention remain active. The philosophical legacy includes attention to public criteria, dispositions, learning, and the distinction between intelligence and a hidden duplicate of intelligent activity. Identity theory and functionalism arose partly in response to behaviorism, not as inevitable stages of a single unanimously accepted progression.

### Further Reading


- Ryle, Gilbert. 1949. *The Concept of Mind*.
- Graham, George. "Behaviorism." *Stanford Encyclopedia of Philosophy*. Revised January 13, 2023.
- Smart, J. J. C. 1959. "Sensations and Brain Processes." *The Philosophical Review* 68 (2): 141–156.


### Related Writing on This Site

- *Mind, Matter, and Meaning* — Gordon Swobe's book presents the author's argument about mind, consciousness, and meaning.

### My View

Behaviorism was the overcorrection to dualism. It threw out the ghost, and with it the genuine inner states a theory of mind exists to explain. The actor marks the exact spot where the analysis fails: the difference between pretending and hurting is the inside, and behaviorism had defined the inside away. That failure is not a minor gap — it falls on the most basic case the theory needed to handle, the felt quality of the state itself.

What I keep is Ryle's diagnosis. The ghost in the machine really is a bad picture, and my own campaign against reified inner objects is a Rylean move; I recruit his diagnosis while declining his cure. Behaviorism's demand for a public anchor was legitimate too — my answer just runs through world-involving representational relations rather than through behavior. The Chinese Room exposes a behaviorist shortcut that later readers sometimes build into the Turing Test: fluent performance alone does not constitute understanding. Behavior remains defeasible evidence, and the test does not settle what realized organization produces the performance.

The history, as I read it, forms a spine, each position failing in the way that makes the next step necessary:

| Position | Inner states? | Physical? | Where it fails |
|---|---|---|---|
| **Dualism** | Yes | No | No causal story; no science |
| **Behaviorism** | No (analyzes them away) | — | Can't distinguish genuine states from mimicry |
| **Identity Theory** | Yes | Yes (specific substrate) | Multiple realizability |
| **Functionalism** | Yes | Not by definition | Narrow, world-bracketed role leaves content open; wide functionalism may converge with the book's world-involving account |
| **Strong representationalism within a physicalist, world-involving account** | Yes | Yes, without substrate chauvinism | The book's view; embodiment is one powerful realization route, not a universal gate |

**Carbon Chauvinism**

In philosophy of mind, **carbon chauvinism** is a critical label for the assumption that thought or consciousness requires carbon-based biological material simply because known conscious organisms have that composition. The charge concerns an allegedly unjustified restriction, not every scientific hypothesis that particular material properties contribute to consciousness.

The underlying **substrate question** asks what a mental capacity requires of its physical implementation. Does preserving the right causal organization preserve mentality across changes in material, or do some further physical properties matter? Answering that question requires specifying the capacity and the organization at issue. Intelligence, understanding, consciousness, and welfare need not all have the same requirements.

**Multiple Realizability and Functionalism**

Functionalists characterize mental states through what causes them, what they cause, and their relations to other states. If pain occupies a particular causal role, different physical mechanisms might occupy that role in humans, other animals, or artificial systems. This is the proposal of **multiple realizability**, not an empirical demonstration that every alternative material can support pain.

Ned Block's discussion of functionalism identifies two opposing risks. **Chauvinism** excludes genuine minds because the theory demands too much similarity to a familiar realizer. **Liberalism** includes systems that lack minds because its conditions are too permissive. A theory can avoid privileging human biology yet still misclassify a large network of people or a simple machine.

This two-sided problem explains why substrate neutrality is weaker than a complete theory of mentality. Rejecting an arbitrary carbon requirement leaves open which causal, representational, historical, or dynamical conditions a nonbiological system would need.

### Substrate-Sensitive Alternatives

John Searle's Chinese Room argument targets the claim that implementing a program suffices for understanding. Searle holds that thinking systems require relevant causal powers, which brains demonstrably possess. His argument does not restrict those powers to carbon as such: in "Minds, Brains, and Programs" (1980), he allows machines with causal powers equivalent to those of brains. Critics dispute whether the powers he requires could be captured by suitably detailed functional organization.

The simulation analogy frames this dispute. A computer model of digestion does not digest food. A model of a storm does not produce its modeled rainfall. A computationalist replies that the properties constituting cognition may differ from those constituting digestion or rainfall. If cognition depends on implemented information processing, a computer might genuinely realize that processing while simulating the biological details of a brain. Neither analogy alone identifies the relevant properties of consciousness.

Type-identity theories offer a different substrate-sensitive position: a mental kind is identical to a physical kind. Defenders may restrict such identities to a species or system rather than insist that all conceivable pain requires one human neural type.

### Organizational Invariance

David Chalmers's *The Conscious Mind* (1996) argues that, under the actual laws of nature, experience remains invariant when fine-grained causal organization is preserved. This is a claim of natural or nomological invariance, not a claim that the organization entails experience in every metaphysically possible world. In a **fading qualia** scenario, biological neurons are gradually replaced with functionally equivalent components and experience allegedly fades. Because all dispositions to notice and report remain fixed, the subject continues to report ordinary vivid experience. Chalmers argues that such a separation between experience and cognitive responses is implausible.

A **dancing qualia** scenario switches between implementations while preserving organization. If the implementations have different experiences, experience changes without the system being able to notice in the relevant functional sense. Chalmers uses this consequence to support organizational invariance.

The argument is not a deductive demonstration that such changes are contradictory. A critic can question the inference from lack of functional noticing to lack of phenomenal change, or dispute whether the proposed replacements preserve everything relevant. Chalmers himself combines

organizational invariance with a nonreductive theory of consciousness; substrate independence and physicalism are separate commitments.

**Open Questions**

The debate turns on the appropriate level of causal detail, the role of biological dynamics, and how evidence could distinguish competing implementations. Material composition alone does not establish consciousness, but describing two systems with the same abstract program need not establish equivalent physical organization either. A justified substrate-sensitive theory must name relevant properties; a justified substrate-independent theory must explain which preserved organization suffices and why.

**Further Reading**

- Chalmers, David J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.
- Block, Ned. 1978. "Troubles with Functionalism." *Minnesota Studies in the Philosophy of Science* 9: 261–325.
- Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3): 417–424. Accompanying commentary and response continue through page 457.

**Related Writing on This Site**

- *Mind, Matter, and Meaning* — Gordon Swobe's book presents the author's argument about mind, consciousness, and meaning.

**My View**

I reject carbon chauvinism. Consciousness is a natural achievement of physical organization, and nothing in my physicalism privileges carbon as such. The relevant questions concern what fixes a system's contents, which persisting organization bears them, whether some states satisfy the best-supported state-consciousness relation, and what phenomenal profile they present. Embodiment supplies powerful realization routes; practical stake bears separately on autonomy and welfare. None makes carbon a criterion. See *Physicalism* and *Embodiment*.

But rejecting carbon chauvinism settles far less than enthusiasts assume, and Block's two-sided warning remains right: the cure for chauvinism is not liberalism. Fluency and manufacture settle nothing by themselves. Training can give text-model mechanisms causally internalized derived content, corpus-mediated history may transmit public reference, and deployed agents may display domain-specific competence or even partial understanding. Evidence for broader, world-involving integration and phenomenal consciousness remains architecture- and subject-sensitive. A silicon system with the relevant history and organization remains eligible; a system made of neurons would receive no automatic pass. The substrate is not the point in either direction.

**The Chinese Room**

The Chinese Room is John Searle's thought experiment against **strong AI**, as he defines it: the claim that an appropriately programmed computer literally understands and that its program explains the relevant human cognitive capacity. Introduced in "Minds, Brains, and Programs" (1980), the argument distinguishes successful formal processing from understanding what symbols mean.

Searle contrasts strong AI with **weak AI**, the use of computers to model and investigate cognition. The argument does not dispute the usefulness of such models. Nor does it claim that no machine could think: Searle regards brains as machines and allows artificial systems with the requisite causal powers.

### The Setup and Inference

An English speaker who knows no Chinese receives Chinese inscriptions and follows English instructions for manipulating them by shape. The rules produce replies indistinguishable, to Chinese-speaking readers, from appropriate answers to questions. The operator nevertheless understands none of the Chinese.

Searle argues that implementing a program cannot therefore suffice for understanding. The operator has the formal operations a program supplies while lacking the semantics, or meaning, of the Chinese material. His broader conclusion holds that no purely formal program, however elaborate, is by itself sufficient for intentionality.

Critics distinguish two claims: the operator does not understand, and the implemented system does not understand. The thought experiment makes the first vivid. Whether it establishes the second is the central dispute. Public Chinese inscriptions also retain their conventional meaning even when an operator cannot interpret them; the issue is understanding by the proposed cognitive system, not whether the inscriptions have any meaning at all.

### The Systems Reply

The **systems reply** attributes understanding to the operator, rules, memory, and processing organization together. Individual neurons do not understand Chinese either; understanding need not belong to each component of a system that understands.

Searle responds by having the operator memorize the rules and perform all the operations internally. The operator still understands no Chinese, he argues, even after incorporating the whole system. Defenders of the reply maintain that introspective ignorance at the operator level does not establish the absence of a distinct implemented organization with Chinese-language capacities. The parties disagree about where understanding would be attributed and what evidence the imagined case provides about it.

### The Robot Reply

The **robot reply** adds sensors and action. A robot can connect language with objects it encounters and actions it performs. Its symbols would participate in more than relations to other symbols.

Searle argues that adding inputs and outputs changes nothing essential if the operator continues to receive uninterpreted marks and execute formal instructions. He also observes that invoking world-related causal powers goes beyond the thesis that a program by itself suffices.

Critics reply that the complete robot's causal relationships, learning, and coordinated activity may establish capacities that the operator's limited perspective cannot reveal. The disagreement concerns whether grounding and organization add genuine understanding, and whether their contribution can itself be computationally described.

### The Brain-Simulator Reply

A further reply proposes reproducing the detailed processing of a Chinese speaker's brain. Searle argues that a simulation need not reproduce the causal powers of what it models, just as simulated digestion does not digest. Computationalists accept the general simulation/realization distinction but dispute its application: a simulation of neural activity might simultaneously realize the information-processing organization that constitutes understanding.

Consequently, the example must specify which properties are preserved. Merely labeling a system a simulation settles neither that it understands nor that it cannot.

### Scope and Contemporary AI

The Room remains relevant to distinctions between fluent performance, internal representation, and understanding. It does not function as an empirical test of every artificial architecture. Hand-coded rules, learned neural networks, and deployed agents with memory, sensors, tools, and feedback differ in ways that rival theories may regard as decisive.

Supporters regard the case as showing that syntax never suffices for semantics. Opponents argue that it assumes, or fails to establish, the absence of system-level understanding. Some intermediate positions accept that a purely formal description leaves meaning underdetermined while allowing an implemented system's history and use to supply it.

These are conclusions drawn from the argument, not a consensus verdict delivered by it. The continuing questions concern the nature of computation, what fixes representational content, and how to identify a system that genuinely uses that content in understanding.

### Further Reading


- Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3): 417–424, with commentary and author response through page 457.
- Copeland, B. Jack. 1996. "What Is Computation?" *Synthese* 108: 335–359.
- Mallory, Fintan. 2026. "Teleosemantics for Neural Word Embeddings." *Mind & Language*. Online publication. DOI: 10.1111/mila.70037.


### Related Writing on This Site

- *Mind, Matter, and Meaning* — Gordon Swobe's book presents the author's argument about mind, consciousness, and meaning.

### My View

I recruit Searle heavily on the syntax/semantics gap; the Room remains among the cleanest arguments that formal operations considered alone do not fix meaning or identify an understander. Its strongest lesson concerns insufficiency, not impossibility. Public symbols, a rulebook, and fluent output do not establish that one system integrates content across a cognitive life.

Where I part company is Searle's **biological naturalism**. Searle grounds consciousness in "the right kind of biology," the causal powers of neurons. I ground content in causal and learning history and attribute understanding at the level where that content enters persistent, flexible, world-answerable organization. Such organization may run through computation and need not

run on neurons. For a bare text-only model, integrated, world-involving understanding remains unestablished; for a larger system with memory, tools, perception, and action, the evidence must be examined rather than inherited from the Room. Artificial consciousness remains open in principle, and fluency alone settles none of it.

### Computation and the Church-Turing Thesis

The **Church-Turing thesis** identifies functions calculable by an effective procedure with functions computable by a Turing machine. An effective procedure consists of determinate instructions that can be followed mechanically, step by step, without ingenuity. Alonzo Church and Alan Turing proposed different formal treatments of calculability in the 1930s; their treatments, along with other major formalisms, agree in the functions they classify as computable.

The thesis concerns the relationship between an initially informal idea and a mathematical characterization. The equivalence of particular formal systems can be proved mathematically. Identifying their common reach with everything effectively calculable requires an analysis of effective calculation; it is not simply another equivalence theorem.

### Effective Calculation and Its Limits

A Turing machine is an abstract device with a finite control, a read-write head, and an unbounded tape. It follows rules specifying its next action from its current state and the symbol it reads. A universal Turing machine can simulate any other Turing machine when supplied with an appropriate description.

Universality does not mean that every mathematical question has a computational solution. Some functions are not Turing-computable. Nor does it mean that every computable problem can be solved with feasible time or memory. **Computability** asks whether an effective method exists; **complexity** asks what resources an algorithm requires.

No ordinary manufactured computer possesses an actually infinite tape. The mathematical model describes what finite rule-governed procedures can do when arbitrary finite amounts of storage are allowed, rather than attributing unlimited physical resources to a particular machine.

### Physical and Complexity Theses

A **physical Church-Turing thesis** extends the claim to physically realizable computation. There are several formulations, differing over which physical systems, measurements, and resources they admit. Paolo Cotogno's survey of hypercomputation examines proposals involving analog systems, infinite processes, and other nonstandard resources. Their mathematical specification does not establish their physical feasibility.

An **extended Church-Turing thesis**, in a common complexity-theoretic sense, concerns efficient simulation rather than bare computability. Quantum computation challenges familiar claims about efficient classical simulation without thereby establishing computation of functions beyond the Turing-computable class. Computational speed and computational reach are different questions.

The original thesis about effective procedures should therefore not be cited as if it already proved every stronger claim about physics.

### Computational Theories of Mind

**Computationalism** proposes that cognition consists in, or depends centrally on, computation. **Computational functionalism** specifies mental roles as computational states within a broader causal-role theory. Neither follows merely from the Church-Turing thesis. A process can be mathematically modeled without its explanatory nature having been established as computational.

Supporters point to computation's ability to explain complex reasoning through simple local operations. It offers a mechanism without requiring an inner agent already capable of the whole task. They also argue that appropriately implemented organization may be realized in different materials.

Critics ask whether formal processing accounts for meaning, consciousness, and environmental relations. Searle's Chinese Room argues that program implementation does not suffice for understanding. Its opponents maintain that understanding may belong to the whole implemented system rather than to a component executing individual instructions.

### Implementation, Simulation, and Meaning

A physical computation needs an account of how physical states and transitions implement formal ones. If any arbitrarily chosen correspondence counted, almost any complicated object could be described as executing any finite computation. B. Jack Copeland's "What Is Computation?" (1996) addresses this problem by distinguishing legitimate implementation relationships from nonstandard mappings. Different theories emphasize causal structure, architecture, counterfactual behavior, or mechanistic function.

Implementation also differs from simulation. Computing a description of a storm does not produce its rainfall. Whether computing a model of cognition realizes cognitive capacities depends on which physical or organizational properties those capacities require. The storm analogy illustrates a distinction but does not settle that further question.

Finally, a computational description and a semantic interpretation answer different questions. Some theories ground representation through learning, selection, or causal relations to an environment; others give formal organization a larger role. The Church-Turing thesis does not decide between them. It provides a framework for effective computation, not a standalone theory of mind.

### Further Reading


- Copeland, B. Jack. 1996. "What Is Computation?" *Synthese* 108: 335–359.
- Cotogno, Paolo. 2003. "Hypercomputation and the Physical Church-Turing Thesis." *The British Journal for the Philosophy of Science* 54: 181–223.
- Copeland, B. Jack. "The Church-Turing Thesis." *Stanford Encyclopedia of Philosophy*. Revised December 18, 2023.

**Related Writing on This Site**

- *Mind, Matter, and Meaning* — Gordon Swobe's book presents the author's argument about mind, consciousness, and meaning.

**My View**

The inference from "the brain is computable" to "a computer running the right program would understand" loses each of its joints once the three claims are kept separate. Computationalism does not follow from the CTT. And even granting computationalism for argument's sake, an abstract program or formal symbol manipulation considered by itself does not constitute meaning. A running implementation may realize content-fixing history, producer-consumer use, embodiment, and integration that the abstract program leaves unspecified; that wider possibility must be judged on its actual organization.

I don't deny that brain processes are, in the relevant sense, computable. I deny that the concession establishes program sufficiency, because the inference from computability to understanding runs through three joints: **CTT ≠ computationalism** (computability of brain processes does not entail that minds are computers); **simulation ≠ realization** (calling something a simulation does not settle which relevant powers its physical implementation realizes); and **syntax ≠ semantics** (formal operations, taken by themselves, do not constitute meaning). The Chinese Room pressures the third joint at the level of the operator and the bare formal description; it does not independently settle what a suitably organized implemented whole could understand. That is why Havlík's challenge deserves an answer rather than a dismissal, and why the wider case turns on the implementation's content-fixing relations and coordinated capacities.

Understanding requires content-guided capacities to constrain one another within an answerable system; it does not require every content involved to be original. Formal structure alone supplies neither content-fixing history nor that integration. A computational implementation may realize both, and its actual history and organization must decide.

**Derived and Original Intentionality**

The distinction between **derived** and **original intentionality** concerns how something acquires representational content: what makes a state or inscription about something and gives it conditions of correctness. In John Searle's terminology, derived intentionality depends on the intentions and interpretive practices of minded users; original or intrinsic intentionality belongs to mental states without acquiring its meaning solely through such interpretation.

A written instruction to *stop* means what it does through a linguistic and social practice. Searle contrasts this with a person's intention to stop, which he treats as genuinely intentional rather than meaningful only because someone else interprets it. Whether this contrast identifies a fundamental difference, and how to explain the original side, remain disputed.

**What "Original" Does Not Mean**

Original intentionality does not mean unprecedented output. A novelist and a language model can both produce a sentence never previously written. The philosophical question asks what fixes

its meaning and what, if anything, the producing system represents. A novel inscription can still derive its meaning from an established language.

Original also need not mean uncaused, innate, or unaffected by social learning. Human concepts develop through interaction with other speakers. A naturalistic theory can appeal to evolutionary or learning history while asking whether the resulting content depends on someone interpreting the state. Causal inheritance alone neither demonstrates nor excludes semantic independence.

"Derived" does not mean meaningless or causally inert. A map can guide travel; a learned internal representation can guide subsequent processing. The dispute concerns the basis of its content, not whether it makes a difference to behavior.

### Competing Accounts

Searle links original intentionality to mental capacities produced by the causal powers of brains. His Chinese Room argues that executing formal rules is insufficient for the relevant understanding. Critics contend that his analysis gives the operator's perspective too much authority over the capacities of the implemented whole.

**Teleosemantic accounts** explain content through the functions of mechanisms and their histories of selection or learning. A producer generates a state; a consumer uses it in a way whose success depends on particular environmental conditions. Those conditions can supply a standard for correct and incorrect representation. The challenge is to show that a proposed mapping explains the mechanism rather than merely reflecting an observer's preferred description.

**Interpretationist approaches** give a larger role to successful attribution of beliefs and desires. They question whether content must possess an interpretation-independent origin to count as genuine. Other naturalistic approaches appeal to causal tracking, informational relations, or inferential organization. These positions disagree about both the distinction's importance and its application.

### Mechanism Content and Whole-System Capacities

Three questions should be kept apart:

1. What, if anything, fixes the content of a particular state?
2. At what level should that content be attributed: a component, a coupled mechanism, or a whole agent?
3. What capacities does the larger system possess: classification, inference, understanding, agency, or conscious experience?

A subpersonal representation need not constitute a miniature subject. Conversely, a system containing meaningful states does not automatically possess every cognitive capacity associated with a person. Content, understanding, and consciousness require their own accounts, even when a theory connects them closely.

### Neural Word Embeddings

Fintan Mallory's "Teleosemantics for Neural Word Embeddings" (2026) applies a consumer-based account to word2vec. It assigns embeddings content concerning actual word-types and

their contexts, rather than simply treating numerical proximity as meaning. His proposal joins a descriptive aspect, concerning a word's context, with a directive aspect guiding downstream context prediction.

This gives a concrete alternative to the claim that all artificial content is merely imposed from outside. But the proposed content has a limited target. Tracking facts about use of the word *maple* differs from representing the tree outside a window. A test that changes the tree while holding linguistic data fixed does not alone refute a claim about linguistic distributions.

An external dictionary or engineered training objective raises questions about semantic dependence. Neither is a quick verdict: human learners also inherit public standards. The relevant dispute concerns whether training and subsequent use establish functions with correctness conditions, and whether those conditions remain exhausted by the originating interpreting practice.

Mallory's analysis of embeddings also does not establish whole-model understanding or consciousness. Those require further arguments about the containing system. The distinction remains useful only when its application names a specific content, level, and explanation, rather than classifying every human or artificial state at once.

**Further Reading**

- Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3): 417–424, with commentary and response through page 457.
- Mallory, Fintan. 2026. "Teleosemantics for Neural Word Embeddings." *Mind & Language*. Online publication. DOI: 10.1111/mila.70037.

**Related Writing on This Site**

- *Mind, Matter, and Meaning* — Gordon Swobe's book presents the author's argument about mind, consciousness, and meaning.

**My View**

I retain the original/derived distinction as a diagnostic of semantic dependence, which can vary by content and grain. Selection, learning, and artificial training may establish original content where history and producer-consumer use fix accuracy conditions not exhausted by an interpreting practice. An objective causal role alone does not settle that question; neither does the ancestry of the vocabulary, corpus, task, or loss. Mallory supplies a serious account of mechanism-level content about actual word-types and linguistic distributions. Whether that content is original at this limited grain remains unsettled. Either kind of content leaves open whether the whole system believes, understands, perceives, or experiences anything.

This gives internal states their fair due. They can carry derived content without being passive inscriptions, miniature subjects, or proof of understanding by the machine that contains them. Mechanism-level names where the content operates; it does not name a third semantic grade between derived and original.

### Direct Realism

Direct realism holds that ordinary successful perception makes us aware of mind-independent objects themselves, rather than only of mental intermediaries from which their existence must be inferred. Looking at an apple ordinarily involves seeing the apple, not first seeing a private proxy and then concluding that an apple caused it.

"Direct" does not mean causally unmediated. Light, sensory organs, and neural processing enable perception. The philosophical issue concerns whether awareness of a distinct intermediary mediates awareness of the object, not whether perception has a physical mechanism.

### Varieties of Directness

**Naive realism** gives a relational account of successful experience: the perceived object and its properties help constitute the experience itself. The apple does not merely cause a self-contained experiential state; the subject's relation to it belongs to what that experience fundamentally is. Many naive realists adopt metaphysical disjunctivism to distinguish this case from hallucination.

**Representational direct realism** allows experience to have content specifying how the world appears. In successful perception, the object represented is the object perceived; representing need not involve perceiving a representation. Such accounts can allow perception and hallucination to share representational features while differing in accuracy and appropriate connection to an object.

These accounts need not agree on the exact relationship between content and phenomenal character, or on what makes awareness direct. Direct realism, intentionalism, and a transparency thesis therefore name different commitments. None should be substituted for the others without explanation.

### Illusion and Perceptual Variation

A straight stick can look bent in water. A traditional argument concludes that something bent must be immediately experienced, and that this cannot be the straight physical stick. A representationalist rejects that inference: the stick can be represented inaccurately without there being a second, bent object.

A naive realist can instead explain the appearance through the actual object, viewing conditions, and the way their relation affects the perceiver. Accounts differ over how much illusion is a misrepresentation and how much it concerns resemblance, perspective, or a misleading judgment about what is seen. They do not all treat illusion as equivalent to hallucination.

Likewise, a round coin presents an elliptical outline from an angle. That outline need not imply either a false representation of the coin's shape or awareness of a private ellipse. A theory must distinguish perceived shape, perspectival appearance, and judgment about shape. Treating every variation as an error prejudges this analysis.

### Hallucination and the Common-Kind Question

A hallucination can be indiscriminable, to the subject, from seeing an object. Yet no appropriate object is perceived. Critics argue that the common subjective character reveals a shared experiential nature, and that the real object cannot therefore constitute that nature.

Disjunctivists reject the inference from indiscriminability to one fundamental kind. A subject may be unable to tell which state they occupy without those states being metaphysically identical. Representationalists instead offer a common account of how the world is presented, alongside additional conditions distinguishing seeing from hallucinating.

The disagreement concerns explanatory priorities. A relational theory emphasizes what successful perception involves. A common-factor theory emphasizes similarities in subjective character, judgment, and behavior across successful and unsuccessful cases. Neither description alone resolves all the causal and phenomenal questions.

### Causation, Time, and Neural Processing

Light reaches a perceiver after a delay. Seeing a distant star consequently depends on its earlier condition. A direct realist can understand this as awareness of a temporally distant object or event; directness need not require simultaneous existence of the object as it appears. Cases involving vanished objects remain a test of the theory's account of perceptual relations.

Neural mediation also does not alone establish that neurons are the objects of vision. A distinction between a representational vehicle and what it represents helps explain why a brain-based state could concern an external object. But that distinction supplies no automatic account of successful perception. The right causal connection must still distinguish seeing an object from having a matching experience accidentally caused by it.

### Principal Disputes

Relationalists question whether object-independent content adequately explains the object's presence in experience. Representationalists question whether relational accounts adequately explain hallucination and shared effects without a common experiential factor. Both face questions about perceptual error, attention, perspective, and deviant causal chains.

Transparency, the observation that attending to experience often directs attention to apparent worldly features, supports several interpretations. It does not independently prove that experience has no intrinsic features or that one particular direct-realist theory is correct.

### Further Reading


- Martin, M. G. F. 2004. "The Limits of Self-Awareness." *Philosophical Studies* 120: 37–89.
- Crane, Tim, and Craig French. "The Problem of Perception." *Stanford Encyclopedia of Philosophy*. Revised August 18, 2021.
- Soteriou, Matthew. "The Disjunctive Theory of Perception." *Stanford Encyclopedia of Philosophy*. Revised July 28, 2026.

**Related Writing on This Site**

- *Mind, Matter, and Meaning* — Gordon Swobe's book presents the author's argument about mind, consciousness, and meaning.

**My View**

I defend intentionalist direct realism. Representational content is not a thing we perceive; it is a *way* we perceive. That single move answers all five classical objections (illusion, hallucination, perceptual variation, time-lag, neuroscience) without retreat to sense-data — and it means representation and directness were never opponents. Representation is what makes directness possible for finite, fallible, perspectival perceivers.

The view threads between two unstable extremes. Naïve realism and disjunctivism preserve directness by making objects constituents of experience, but they buy it at too high a price: hallucination's genuine phenomenal character goes unexplained, and even veridical misattribution becomes awkward. Indirect realism keeps a unified account of perception and hallucination but sacrifices directness and generates global skepticism. The intentionalist route: misrepresentation replaces the sense-datum. We do not have to choose between being honest about the mind and being direct about the world.

These commitments support one another but answer different questions. Intentionalist direct realism concerns the objects and structure of perception; strong representationalism concerns phenomenal character; transparency supplies phenomenological evidence against an inner intermediary. Disjunctivism is also a direct realism, which is why "direct realism" alone does not name my view.

**Disjunctivism**

Disjunctivism names a family of views about the differences between successful perception and perceptual error. **Metaphysical disjunctivism** denies that genuine perception and a subjectively matching hallucination share the same fundamental experiential nature. **Epistemological disjunctivism** holds that successful perception can provide reasons or warrant unavailable in an indistinguishable unsuccessful case. The theses are related but distinct.

The term derives from a disjunction: someone either genuinely sees an object or undergoes an experience indistinguishable from seeing it. The point goes beyond this uncontroversial classification. A metaphysical disjunctivist denies that a single underlying experiential kind explains both alternatives, while an epistemological disjunctivist denies that their subjective similarity limits their warrant to one common factor.

**The Common-Kind Assumption**

The argument from hallucination begins with a possible experience that the subject cannot distinguish from seeing an ordinary object, although no such object is perceived. A **common-kind theory** proposes a fundamental experiential state shared across the good and bad cases. The actual object then distinguishes perception from hallucination without constituting that shared nature.

Sense-datum theories and many representational theories accept versions of this proposal, but identify the common element differently. Disjunctivists argue that subjective indiscriminability does not entail metaphysical identity. Not knowing whether one is perceiving does not establish what kind of state one occupies.

J. M. Hinton's early formulation challenges the assumption that reports of appearances must report a positive experiential occurrence common to both cases. Later accounts develop this challenge into more explicit theories of experience and perceptual knowledge.

### Naive Realism and Hallucination

Many metaphysical disjunctivists defend **naive realism**: in successful perception, mind-independent objects and their properties partly constitute the experience. M. G. F. Martin's "The Limits of Self-Awareness" (2004) develops a defense that emphasizes the limits of introspection. An object-involving experience need not reveal, merely through introspection, that it has that object-involving nature.

Hallucination requires a different explanation because no corresponding perceptual relation obtains. A **negative account** characterizes the relevant hallucination through its indiscriminability from perception, rather than through a positive phenomenal kind shared with perception. This is a claim about how to characterize the experience, not a denial that hallucination occurs or affects thought and action.

Other approaches give hallucination a more positive account while preserving some disjunctive commitments. Disjunctivism need not deny all common features, neural causes, descriptions, or effects. Its stronger formulations deny a shared *fundamental* experiential nature. Disjunctivists also differ over illusion: some treat it as genuine perception of an object under misleading conditions rather than grouping it with objectless hallucination.

### Epistemological Disjunctivism

John McDowell's epistemological use of the approach rejects the idea that the best perceptual reason must be available equally in a hallucination. In the good case, seeing that something is so can place the fact itself within the subject's rational access. In the bad case, only an appearance of that access occurs.

A critic asks how this stronger reason can be available if the subject cannot distinguish the cases. Defenders distinguish having a factive reason from having an independent procedure that guarantees recognition of every good and bad case. Accounts differ over the role of reflection and self-knowledge. A commitment to stronger warrant in successful perception does not, by itself, require a complete metaphysical disjunctivism about phenomenal character.

### Objections and Replies

**Shared causal effects.** Perception and matching hallucination can produce similar judgments, memories, and behavior. A common-kind theorist argues that a shared experience offers a unified explanation. Disjunctivists reply that common effects need not establish common fundamental causes; however, the theory still needs an account of the similarities.

**The causal argument.** Appropriate stimulation of the perceptual system can allegedly reproduce experience without an object. Critics infer that the same inner cause must produce the same experiential kind when an object is present. Disjunctivists deny that matching proximal causes settle the nature of a relational event. The dispute concerns whether the common inner cause screens off the explanatory contribution of the actual perceptual relation.

**Hallucinatory character.** A negative description may seem insufficient to explain what a hallucination positively feels like. Defenders argue that indiscriminability captures limits of self-awareness without introducing an unsupported common kind. Critics ask whether that dependent characterization adequately explains subjective reasons and phenomenal specificity.

The debate remains a disagreement over explanatory structure, not a choice between acknowledging hallucination and denying it.

### Further Reading

- Martin, M. G. F. 2004. "The Limits of Self-Awareness." *Philosophical Studies* 120: 37–89.
- Soteriou, Matthew. "The Disjunctive Theory of Perception." *Stanford Encyclopedia of Philosophy*. Revised July 28, 2026.
- Crane, Tim, and Craig French. "The Problem of Perception." *Stanford Encyclopedia of Philosophy*. Revised August 18, 2021.

### Related Writing on This Site

- *Mind, Matter, and Meaning* — Gordon Swobe's book presents the author's argument about mind, consciousness, and meaning.

### My View

Disjunctivism is the most respectable rival to my intentionalist direct realism in the philosophy of perception — the one direct-realist competitor that has to be engaged rather than dismissed. Both disjunctivism and intentionalist direct realism refuse the veil of sense-data; both insist that in ordinary perception the world itself is present to the mind. The disagreement is over *how* to secure that result.

The disjunctivist secures direct realism by denying that genuine perception and a subjectively indistinguishable hallucination share one fundamental experiential nature. My view secures direct realism a different way — through representational content. I treat common content as an abductive proposal, not as something indiscriminability hands us: it earns support by explaining hallucination's positive character and the matched rational role while current satisfaction and coupling explain the difference between seeing and merely seeming to see.

The shared motivation matters. Disjunctivism and intentionalist direct realism are not on opposite sides of the veil. Where McDowell uses disjunctivism to keep the world in view, my view uses transparency and externalist content to the same end. The quarrel is intramural: two anti-Cartesian direct realisms disagreeing over mechanism. See *Vehicle and Content*, *Transparency*, and *John McDowell*.

### Dualism

Dualism in philosophy of mind holds that the mental and the physical are fundamentally different in some respect. The strongest familiar version distinguishes mental and physical substances; other versions distinguish irreducible properties within one world. Dualism should not be equated with every claim that psychological explanations or mental concepts differ from physical ones. Physicalists can accept such differences while denying a further nonphysical reality.

The principal motivation is that consciousness, intentionality, agency, or personal identity seems difficult to accommodate within an exclusively physical ontology. Which phenomenon supplies the difficulty, and what sort of distinction follows, varies across dualist theories.

### Substance and Property Dualism

**Substance dualism** treats subjects of mental life as substances distinct from physical bodies. René Descartes's account contrasts thinking substance with extended substance. His view nevertheless makes human life a union of mind and body, not merely a mind observing a body from afar. Contemporary substance dualism has defenders, although it does not dominate current philosophy of mind.

**Property dualism**, in the ontological sense used here, holds that at least some mental properties are nonphysical: they are not wholly constituted by the physical base. This goes beyond saying that mental vocabulary or explanations cannot be reduced to physical ones. It need not introduce an immaterial subject alongside the organism. Consciousness may instead be a fundamental property linked to the physical by additional laws, or a strongly emergent property that adds something ontologically new. Ordinary emergence of complex capacities from physical organization does not by itself imply property dualism.

David Chalmers's *The Conscious Mind* (1996) develops **naturalistic dualism**, treating experience as part of nature while resisting its reduction to physical structure and function. Such a view seeks systematic psychophysical laws rather than appealing to violations of natural order. It can also support substrate independence: nonphysical experience might depend on organization rather than on exclusively biological material.

### Arguments for Dualism

**The knowledge argument.** Frank Jackson's original Mary case imagines a scientist who knows all physical facts about color vision but has never seen color. Upon seeing red, she appears to learn something. The dualist argues that, if she learns a previously unknown fact, not every fact was included in the physical account. Physicalist replies distinguish new facts from new abilities, acquaintance, or ways of thinking about facts already known. Jackson later rejected the argument's antiphysicalist conclusion.

**Conceivability arguments.** A philosophical zombie is a physical duplicate of a conscious person lacking experience. Chalmers argues that an appropriately specified conceivability of such a duplicate supports its metaphysical possibility; if possible, consciousness does not follow necessarily from the physical facts. Critics question whether the duplicate has genuinely been

conceived coherently and whether conceivability licenses that modal conclusion. Sophisticated versions require more than ordinary imaginative ease.

**The explanatory gap.** Even detailed accounts of discrimination, report, and neural activity seem to leave unanswered why these processes feel a particular way. Dualists argue that this suggests an irreducible feature requiring additional principles. Physicalists dispute the ontological conclusion: a gap in concepts or explanation need not entail a second kind of property. Not every philosopher who accepts an explanatory gap accepts dualism.

### Mental Causation

An **interactionist** allows mental and physical events to affect one another. The familiar objection asks how a nonphysical cause fits with physical explanations, and whether it requires denying that physical events have sufficient physical causes. Interactionists dispute either the closure premise or the proposed account of causation; the objection must therefore state its physical assumptions rather than treat closure as a definition of nature.

An **epiphenomenalist** allows physical processes to produce experience while denying that experience affects physical events. This avoids one form of competition between causes but faces questions about action, evolutionary relevance, and knowledge of experience. If phenomenal properties do no causal work, how do reports reliably concern them? Defenders offer accounts of knowledge and dependence intended to answer this challenge.

**Parallelism** treats mental and physical sequences as coordinated without direct interaction. Historical versions explain this coordination in different ways. None of these causal options follows merely from calling a theory dualist.

### Continuing Disputes

Dualism must explain the relation between its distinct categories and the dependence of ordinary mental life on physical conditions. Physicalist rivals must explain why experiences appear resistant to physical explanation and how their own identifications preserve the phenomena.

The dispute therefore concerns more than whether current neuroscience has completed an explanation. It concerns the force of knowledge and modal arguments, the interpretation of causal evidence, and what an adequate theory would have to explain. An incomplete physical account does not alone prove dualism; physical correlations do not alone refute it.

### Further Reading


- Chalmers, David J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.
- Chalmers, David J. 2003. "Consciousness and its Place in Nature." In *The Blackwell Guide to Philosophy of Mind*, edited by Stephen P. Stich and Ted A. Warfield.
- Robinson, Howard, and Ralph Weir. "Dualism." *Stanford Encyclopedia of Philosophy*. Revised October 17, 2025.

**Related Writing on This Site**

- *Mind, Matter, and Meaning* — Gordon Swobe's book presents the author's argument about mind, consciousness, and meaning.

**My View**

The epistemic gap is real but lodged in us and our concepts, not in the world (see *The Hard Problem of Consciousness*). The inference from "we can't explain how" to "it must be non-physical" is the unearned step that does all the dualist's work. Two moves block it: the transparency of experience and the identity claim deflate the qualia-as-intrinsic-inner-properties picture the zombie and knowledge arguments require; and the phenomenal concept strategy explains *why* the gap should feel exactly this unbridgeable even if physicalism is true. Frank Jackson later abandoned the knowledge argument and became a physicalist, chiefly because epiphenomenalism made our knowledge of phenomenal properties self-stultifying; his mature route differs from this book's transparency, identity, and phenomenal-concept strategy. See *Frank Jackson*.

My own position (physicalism, reductive but non-eliminative and relational) is offered not as a denial of the dualist's data but as a better home for it. The felt character of experience is preserved, as world-presenting representational content, without a second substance, a second kind of property, or irreducible psycho-physical laws. See *Identity Theory* and *Intentionalism*.

**Embodiment**

Embodiment concerns the ways bodily organization and interaction with an environment shape cognition. Perception, skilled movement, emotional regulation, and concepts may depend on what kind of body an agent has and how it uses that body. The term names a family of research programs rather than one universally accepted requirement for having a mind.

A claim that bodily activity influences cognition differs from a claim that it partly constitutes cognition. It also differs from a claim that no disembodied system could ever think or experience. Accounts of embodiment range from experimentally modest hypotheses about particular tasks to broad alternatives to traditional computational models.

**Sources and Motivations**

Phenomenological approaches emphasize the **lived body**: the body as the center of practical orientation and perceptual engagement, not merely an object inspected from outside. Merleau-Ponty's work motivates analysis of habitual skill and the ways a world presents possibilities for action.

Francisco Varela, Evan Thompson, and Eleanor Rosch's *The Embodied Mind* (1991) develops an **enactive** approach. It treats cognition through the history of an organism's coupling with its environment, challenging the picture of an independently specified world simply copied into an independently specified mind. Later enactive approaches differ over how much representation and computation they retain.

Cognitive-scientific motivations include reducing explanatory dependence on internal representations where bodily dynamics and environmental structure already explain performance. The

body's morphology can simplify a control problem: a controller need not calculate every detail that the shape and material of a moving system help determine.

### Related but Distinct Programs

**Embodied cognition** gives bodily properties a substantive role in cognitive explanation. **Embedded cognition** emphasizes the environment in which cognition operates. **Extended cognition** proposes that, under suitable conditions, processes beyond the organism partly constitute a cognitive system. **Enactive cognition** emphasizes active engagement and, in some versions, autonomous self-maintenance.

These claims can be combined, but accepting one does not establish all the others. A theorist may recognize that an external notebook helps memory while denying that it belongs to the memory system. Another may accept external cognitive components without requiring cognition to reject internal representations.

A similar distinction applies to semantic externalism. That content depends partly on environmental or social relations does not entail that every episode of thinking currently extends beyond the brain. Historical content dependence and present process constitution are separate theses.

### Illustrative Cases

A skilled cyclist adjusts balance through ongoing movement and feedback. Embodied accounts ask whether a complete explanation must model the coupled rider-bicycle-environment dynamics rather than attributing control entirely to an internal sequence of calculations. Computational accounts can reply that sensorimotor organization belongs within the implemented computational system. The dispute concerns explanatory adequacy, not the mere existence of feedback.

A brain-in-a-vat case tests a different claim. Suppose a brain receives inputs matching those of an embodied subject. An internalist may assign the same relevant mental states to both. An externalist asks how their histories and environments fix reference. A brain newly removed from ordinary life differs from one created in a vat with no such history. The case does not have one verdict shared by every embodiment theory, nor does a reference argument alone settle phenomenal consciousness.

### Objections and Replies

**Causal versus constitutive involvement.** Bodily movement can alter a thought without being part of the thought. Critics argue that evidence of influence is often overstated as evidence of constitution. Defenders seek task-specific accounts showing why isolating an internal mechanism loses the organization that explains the capacity.

**Abstract and offline thought.** Mathematics, planning, dreaming, and thought about absent objects seem possible without current sensorimotor engagement. Moderate accounts appeal to capacities acquired through earlier engagement; stronger approaches must explain how their analysis extends beyond ongoing action. Invoking historical embodiment cannot substitute for specifying the actual mechanism.

**Compatibility with computation.** Some embodied approaches supplement computational models, while others reject them. Critics question whether the proposed alternative explains the same range of capacities. Defenders reply that an approach centered on bodily dynamics can reveal mechanisms obscured by a purely symbolic model. Neither side can settle the dispute by definition.

### Artificial Systems and Open Questions

Robots provide concrete tests of how sensors, action, learning, and morphology contribute to competence. Text-trained systems also have learning histories and connections to public language, although these differ from direct bodily engagement. Adding a body does not automatically establish understanding, consciousness, or welfare; lacking one does not, without further premises, establish their impossibility.

The most informative questions specify the capacity at issue and the contribution the body allegedly makes: enabling, shaping, constituting, or supplying a history. This preserves embodiment's explanatory possibilities without turning it into one all-purpose threshold.

### Further Reading

- Varela, Francisco J., Evan Thompson, and Eleanor Rosch. 1991. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press.
- Shapiro, Lawrence, and Shannon Spaulding. 2021. "Embodied Cognition." *Stanford Encyclopedia of Philosophy*.
- Dreyfus, Hubert L. 2007. "Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian." *Philosophical Psychology*.

### Related Writing on This Site

- *Mind, Matter, and Meaning* — Gordon Swobe's book presents the author's argument about mind, consciousness, and meaning.

### My View

My position: embodiment is not a mystical requirement and not a single threshold. It is a cluster of causal-organizational achievements. Sensorimotor coupling matters especially to perceptual mode and world-guided action; bodily regulation and self-maintenance matter to affect, autonomy, stake, and welfare; causal history and producer-consumer organization matter to content. Biology monopolizes none of them.

Current text-only LLM mechanisms may carry causally internalized derived content, and corpus-mediated causal history may transmit some public reference. That grant does not establish whole-system understanding or feeling. The harder questions concern whether content-bearing states enter one unified, persistent, world-sensitive economy of memory, correction, planning, and action, and whether any realized mode meets the book's account of consciousness. Embodied robotic systems offer one powerful route to those achievements, not the only conceptually possible route. Stake remains important because a system with welfare deserves a different kind of concern; it does not decide content or consciousness in advance.

### Functionalism

Functionalism characterizes mental states through their causal roles: the inputs that produce them, the behavior and other states they produce, and their relations to other mental states. Pain, for example, might involve a state commonly caused by bodily damage that interacts with beliefs and desires and tends to produce protective action. The actual role must allow exceptions: pain can occur without damage, and a person can suppress its expression.

Unlike simple behaviorist analyses, functionalism permits internal states that help cause behavior. Unlike an unrestricted type-identity theory, it need not identify a mental kind with one specific neural kind. Its central claim concerns organization, not an exclusively biological material.

### Multiple Realizability

Hilary Putnam's early machine-functionalism helped develop the possibility that different physical systems realize the same mental kind. An alien could share a mental capacity with humans without sharing their neural anatomy, provided it realized the relevant organization. This **multiple realizability** motivates functionalism but does not by itself prove it: other theories can permit different realizers too.

A functional **role** also differs from its **realizer**. The role specifies causal relationships; the realizer is the state or mechanism that actually occupies it in a given system. Functionalists disagree about whether mental properties should be identified with roles or with occupants of roles under specified conditions.

### Main Variants

**Machine functionalism** models mental organization through machine states and transition rules. **Analytic functionalism** derives roles principally from ordinary concepts and commonsense psychological explanations. **Psychofunctionalism** instead appeals to the organization identified by empirical psychology, which may revise common assumptions.

**Computational functionalism** treats the relevant roles as computational. Functionalism more broadly need not claim that a mind is a program, or that any execution of a matching formal description suffices for mentality. Physical implementation, counterfactual causal structure, and system boundaries require their own specification.

Accounts also differ in scope. A **narrow** role may be specified by internal organization; a **wide** role can include environmental relations and learning history. Externalist objections therefore pressure some forms more directly than others. Putnam's later rejection of internally individuated functional content does not show that every role theory has identical commitments.

### Absent and Inverted Qualia

Ned Block's "Troubles with Functionalism" (1978) asks whether systems satisfying functional descriptions might nevertheless lack experience. In his China-body case, a large population communicates so as to reproduce a proposed functional organization. Many readers doubt that the resulting whole feels anything. Block uses this possibility to challenge the theory's sufficiency conditions, not to establish that all nonbiological systems are unconscious.

Functionalists can reject the intuition, question whether the stipulated system reproduces the full relevant role, or argue that experience belongs to the whole rather than its components. These replies differ: rejecting an intuition is not the same as identifying a missing causal feature. The completeness of the role specification remains central.

The **inverted qualia** objection imagines functionally matching subjects with different color experiences. Critics infer that phenomenal character outruns functional role. Defenders question whether a complete inversion could preserve all relevant discriminations, similarities, memories, and dispositions, or argue that suitably specified representational content excludes the proposed inversion. Whether such replies explain consciousness or build it into the conditions is disputed.

### Meaning and the Chinese Room

Searle's Chinese Room challenges program sufficiency by describing an operator who executes rules without understanding the processed language. The systems reply attributes understanding to the implemented whole. Searle's internalization response and the robot reply generate further disagreement about the role of introspection, grounding, and causal powers.

A theory of causal roles must explain why its states count as beliefs about particular objects or circumstances. Some functionalists incorporate environmental and historical relations; others defend internal or inferential accounts of content. Representation is not a question functionalism can simply ignore, but it need not be a question excluded by every functionalist formulation.

### Strengths and Remaining Questions

Functionalism offers systematic explanations of how beliefs, desires, and perceptions coordinate action. It permits scientific investigation of internal mechanisms and allows material variation where organization is preserved. It also faces Block's opposing risks: excluding genuine minds by specifying roles too narrowly, or including nonminds by specifying them too liberally.

Open disputes concern the required causal grain, the empirical status of multiple realization, environmental dependence, and whether phenomenal consciousness admits a functional analysis. A functional account of one capacity need not settle intelligence, understanding, and consciousness together. Nor does functionalism inherently require either physicalism or a claim that present artificial systems already satisfy its conditions.

### Further Reading


- Block, Ned. 1978. "Troubles with Functionalism." *Minnesota Studies in the Philosophy of Science* 9: 261–325.
- Copeland, B. Jack. 1996. "What Is Computation?" *Synthese* 108: 335–359.
- Chalmers, David J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.


### Related Writing on This Site

- *Mind, Matter, and Meaning* — Gordon Swobe's book presents the author's argument about mind, consciousness, and meaning.

**My View**

Functionalism is the best predecessor to a fully grounded account of mind. It correctly identifies that mental states have causal powers, that material composition alone does not settle mentality, and that organization matters. Computational functionalism adds the proposal that the relevant roles are computational states. More recent work on intrinsic computational functionalism adds a further point worth granting: disciplined causal-computational organization can belong to a physical system independently of an observer's chosen labels. Narrow functionalism still falls short when it treats an internally specified, world-bracketed causal role as sufficient; its computational child inherits that problem when it treats formal state as sufficient.

| What the view captures | What an internal, world-bracketed description leaves open |
|---|---|
| Functionalism: causal roles and counterfactual organization | What the states represent and what fixes their correctness conditions |
| Functionalism: material neutrality and the possibility of multiple realization | Whether different systems actually realize one mental type |
| Computational functionalism: formal computational structure | Whether formal state determines content or every mind-relevant power |
| Wide functionalism: complex world-involving organization | Whether several capacities support partial understanding, and how broadly and persistently they integrate |
| Material neutrality toward nonbiological realizers | Whether any content has the mode and poising required for consciousness |

My objection therefore targets no fully specified physical-functional account merely because someone calls it functionalist. It targets the internal diagram with the world bracketed. Put the relevant history and producer-consumer relations back in, and the states may acquire genuine content. Let several content-guided capacities constrain one another in an answerable system, and partial understanding can begin. Add breadth, persistence, and traffic across perception, memory, inference, correction, planning, and action, and that understanding becomes more integrated and world-involving. A wide functionalist who includes those relations has reached much the same account by another road; the remaining disagreement concerns filing.

The representational repair remains central: phenomenal character consists in a conscious state's complete world-presenting intentional profile, not in bare role or mental paint. For artificial understanding, content marks only the first threshold. Content can guide **domain-specific competence** when it explains flexible inference, correction, learning, planning, or action. **Partial understanding** begins when several such capacities constrain one another within an answerable system. Greater breadth, depth, persistence, and cross-domain traffic make that understanding more integrated and world-involving; they do not create a separately named grade. Current systems offer credible evidence of domain-specific competence, live but subject-sensitive evidence of partial understanding, and fragmentary evidence of broader integration.

Consciousness and autonomous stake remain separate questions. State consciousness requires an independently supported threshold relation; within a conscious state borne by an identified subject, phenomenal character consists in content under a perceptual or bodily-affective mode, with determinacy, field structure, and attentional organization. Reciprocal self-maintenance may ground autonomy, individuality, and practical interest; welfare requires conscious valence. It neither creates content nor constitutes understanding or consciousness. No part of this account privileges neurons, carbon, or biological manufacture.

### Higher-Order Thought Theories

Higher-order theories explain a mental state's being conscious through an appropriate representation of that state. A first-order state concerns the world or body; a higher-order state concerns the subject's being in a mental state. An unconscious perception and a conscious perception can therefore differ because only the latter is represented in the required higher-order way.

**Higher-order thought (HOT) theory** gives this representation a cognitive form. Other higher-order theories describe an inner perceptual monitor or a self-representational structure within a single state. These variants disagree about what kind of awareness is required and how it is organized.

### The Explanatory Target

David Rosenthal distinguishes **creature consciousness**, such as being awake, from **state consciousness**, a mental state's figuring in conscious experience. He also distinguishes being conscious *of* something from a state itself being conscious.

His **transitivity principle** connects the last two: a state is conscious when the subject is conscious of being in it in an appropriate way. This is a theoretical proposal about what makes states conscious, not merely an assertion that conscious people can sometimes introspect.

The proposal permits first-order sensory qualities in states that remain unconscious. The higher-order representation explains their subjective appearance to the person. Different accounts specify differently how first-order qualities and higher-order content determine what an experience feels like.

### Actual Thoughts, Inner Sense, and Dispositions

In Rosenthal's actualist HOT theory, the relevant thought is occurrent, assertoric, and apparently noninferential: it represents the subject as currently being in the state, rather than merely entertaining a possibility or inferring it from external evidence. The thought need not be verbally formulated, deliberately generated, or itself conscious.

Allowing an unconscious HOT avoids an immediate regress. If every consciousness-making thought had to be conscious, a further thought would be needed to make that thought conscious, and so on. Critics can still dispute whether an unconscious representation supplies the required awareness, but a regress does not follow simply from the word *higher-order*.

**Higher-order perception** theories, associated with Armstrong and Lycan, appeal instead to monitoring analogous to perception. Rosenthal argues that the relevant awareness need not

have the sensory character such a model might suggest. Defenders of inner sense dispute this characterization of monitoring.

**Dispositional HOT theories**, associated with Peter Carruthers, appeal to availability for higher-order thought rather than requiring an actual separate HOT on every occasion. Availability can be understood as helping confer higher-order representational content on the first-order state. This is more specific than saying that any state that might later be noticed is already conscious.

**Self-representational theories** integrate awareness of the experience within its own structure. They belong to the wider debate but should not be treated as simply Rosenthal's two-state account under another name.

### Motivations and Evidence

Higher-order theories seek to explain differences between conscious and unconscious processing, apparent immediacy of self-awareness, and the connection between conscious experience and reports about it. Rosenthal also considers cases in which changes in qualitative concepts alter how experience is consciously discriminated.

The interpretation of these phenomena is contested. A first-order theorist may explain finer discrimination through attention or changes in the perceptual state, without a consciousness-making HOT. Correlations with metacognitive performance may reflect reporting or monitoring rather than experience itself. Experiments must distinguish those possibilities; verbal report provides evidence but cannot be assumed to settle the mechanism.

### Mismatch and Targetless Representation

A higher-order representation might describe a first-order state incorrectly. It might even occur without the corresponding first-order target. If a HOT represents the subject as seeing red while the underlying state differs, which content determines the apparent experience?

Rosenthal allows that higher-order misrepresentation can shape subjective appearance, including in cases lacking the expected first-order target. Critics argue that this threatens the claim that a state becomes conscious by being represented: in a targetless case, no such target exists. Defenders distinguish the subject's apparent experience from the actual qualities of the represented state. Other higher-order variants restrict or avoid targetless representations.

Rosenthal's dental-fear example, in which a patient with destroyed dental nerves reports apparent pain amid fear and vibration, illustrates his interpretation of mental appearance, not a clinical demonstration that pain must be generated by a HOT. Expectation and bodily processing offer competing explanations.

### Other Objections and Open Questions

Critics ask whether infants and animals possess the required concepts, whether the account demands overly sophisticated metacognition, and whether self-representation explains phenomenal character rather than only access or report. Defenders offer simpler forms of higher-order representation and argue that sophisticated linguistic self-description is unnecessary.

Higher-order theories can also coexist with accounts of first-order content and its environmental grounding. They are not inherently opposed to every representational theory, nor does representing an internal state automatically commit a theory to a private-object ontology. The central dispute concerns what makes an otherwise unconscious state consciously experienced.

**Further Reading**

- Rosenthal, David M. 2009. "Higher-Order Theories of Consciousness." In *The Oxford Handbook of Philosophy of Mind*. DOI: 10.1093/oxfordhb/9780199262618.003.0014.
- Carruthers, Peter, and Rocco Gennaro. "Higher-Order Theories of Consciousness." *Stanford Encyclopedia of Philosophy*. Revised January 9, 2026.

**Related Writing on This Site**

- *Mind, Matter, and Meaning* — Gordon Swobe's book presents the author's argument about mind, consciousness, and meaning.

**My View**

HOT theories and intentionalism both reject qualia realism — both hold that phenomenal character doesn't consist in intrinsic properties of inner objects. But they diverge crucially:

**HOT: Consciousness requires self-representation**

A state is conscious iff there is a thought about it. Phenomenal character is *relational* — it consists in being represented by a higher-order state. A first-order state with exactly the phenomenal character of a pain could exist unconsciously, with no phenomenal character, simply by lacking a HOT.

**Intentionalism: Phenomenal character consists in world-presenting content**

Among conscious states, phenomenal character consists in content under a perceptual or bodily-affective mode, with the relevant determinacy, field structure, and attentional organization. A separate state-consciousness question asks what makes such content figure in experience rather than remain unconscious processing; poising, recurrence, global availability, and related forms of integration remain competing threshold hypotheses. Intentionalism denies that a higher-order representation is required, not that a consciousness threshold must be explained.

**Three structural advantages of intentionalism over HOT theories:**

1. **No two-state problem**: intentionalism doesn't require a distinct higher-order state. The experience represents the world; its phenomenal character consists in that representational content. One state, one experience.
2. **Transparency explained, not merely accommodated**: the Transparency Thesis falls out naturally from intentionalism — introspection returns us to the world because that's what the experience represents. HOT theories must treat transparency as a datum to be explained after the fact: why, if the HOT is about the first-order state, does introspection seem to reveal the external world rather than an inner state?
3. **Externalism integrated**: intentionalism connects naturally to semantic externalism and teleosemantics — what a state represents depends on its causal-historical relation to the

environment. HOT theories are resolutely internalist: the HOT is about a first-order *mental* state, not about the world. They face the same problems as functionalism regarding content determination.

**The dental fear case reconsidered**: Rosenthal takes dental fear to show that what it is like is determined by the HOT rather than by the first-order state. The intentionalist need not grant that result. Fear, expectation, and vibration may alter the content and bodily-affective mode of the conscious first-order state so that a disturbance is presented as pain. That is misrepresentation within the experience itself. If no suitably organized conscious first-order state occurs, a higher-order thought that one is in pain does not by itself generate pain phenomenology.

### Identity Theory

The mind-brain identity theory identifies mental states or processes with physical states or processes of the brain. An experience and a brain process are, on this view, one occurrence described in different ways, rather than two occurrences joined by a causal law. Identity is therefore stronger than correlation or causal dependence.

The theory gained a prominent modern formulation in U. T. Place's "Is Consciousness a Brain Process?" (1956) and J. J. C. Smart's "Sensations and Brain Processes" (1959). Herbert Feigl developed a related position. These approaches allowed sensations to occur as genuine internal processes while resisting a separate nonphysical domain.

### Empirical Identity and Difference in Meaning

Place and Smart compare mental-physical identification with scientific identifications such as lightning with electrical discharge. Someone can recognize lightning while knowing nothing about electricity. Likewise, recognizing an afterimage need not involve knowing its neural nature.

The identity theorist does not therefore claim that an ordinary sensation word and a neuroscientific expression mean the same thing. Different concepts can refer to the same phenomenon. An objection based only on a speaker's ignorance of neuroscience would not establish that sensation and neural activity must be distinct.

Smart described the proposed identity as contingent and empirically discoverable. Later modal discussions complicate that formulation: an identity discovered empirically may nevertheless hold necessarily if its terms rigidly designate the same entity. The historical terminology should not be silently treated as a resolution of those later questions.

### Type and Token Identity

**Type identity** identifies a mental kind with a physical kind. A strong unrestricted version assigns the same physical kind wherever a particular mental kind occurs. **Token identity** instead identifies each particular mental occurrence with a physical occurrence without requiring every instance of that mental type to share one physical type.

This distinction separates two claims often compressed into "mental states are brain states." A functionalist can permit token-level physical identity while identifying mental types by causal

roles. Some identity theorists also defend identities relative to a species or system, rather than one unrestricted identity across all possible organisms.

"Pain is C-fiber firing" is a familiar classroom illustration, not an established universal neuroscientific identity or a complete statement of every historical identity theorist's position. The substantive theory requires an independently supported account of the relevant physical process.

### Smart's Topic-Neutral Strategy

A report of a yellow-orange afterimage seems to mention a quality not found in a brain process. Smart proposes describing the experience in **topic-neutral** language: something occurs like what occurs when the person sees an illuminated orange under ordinary conditions. Such a description does not decide in advance whether the occurring process is physical or nonphysical.

Smart also distinguishes the experience from an alleged colored internal object. Identifying an experience with a brain process does not require a surgeon to find a yellow-orange patch in the brain. Critics reply that topic-neutral reformulation may still leave the qualitative character unexplained. The strategy blocks certain immediate objections to identity without supplying a complete empirical theory of experience.

Parsimony provides another motivation. If sensations were additional nonphysical events, extra psychophysical laws would be required to connect them to neural activity. Identity theorists argue that identification avoids these additional laws. Opponents maintain that explanatory economy cannot override evidence for irreducible properties.

### Multiple Realizability

Functionalists argue that different organisms, and possibly artificial systems, could share mental states despite different physical implementations. If pain can occur in several physical kinds, an unrestricted identity between pain and one neural kind appears too restrictive.

Identity theorists can question whether the cases instantiate exactly the same mental type, defend more local identities, or argue that the apparent variety conceals a common physical characterization. Whether such moves preserve informative scientific identities is debated. Multiple realizability is consequently a major objection to certain type identities, not an automatic refutation of every mental-physical identity claim.

### Phenomenal and Externalist Objections

Knowledge and conceivability arguments ask whether complete physical knowledge or a physical duplicate could still leave experience out. Identity theorists distinguish concepts from referents and challenge the inference from apparent separability to genuine possibility. Critics contend that these replies do not explain why sensation has its particular felt character.

Semantic externalism supplies another pressure: some beliefs are individuated partly by relations to an environment or community, while a narrowly described brain state does not include those relations. An identity account must specify its scope rather than assume that a successful account of sensations automatically applies to every belief.

The continuing dispute concerns which identities evidence supports, at what explanatory level, and whether physical identification also adequately explains phenomenal character and content.

**Further Reading**

- Smart, J. J. C. 1959. "Sensations and Brain Processes." *The Philosophical Review* 68 (2): 141–156.
- Place, U. T. 1956. "Is Consciousness a Brain Process?" *British Journal of Psychology* 47: 44–50.
- Smart, J. J. C. "The Mind/Brain Identity Theory." *Stanford Encyclopedia of Philosophy.* Revised May 18, 2007.

**Related Writing on This Site**

- *Mind, Matter, and Meaning* — Gordon Swobe's book presents the author's argument about mind, consciousness, and meaning.

**My View**

Identity theory makes three permanent contributions to the dialectic:

1. **The contingent identity model**: mental-physical identity claims can be empirically discovered without being definitionally true. This applies equally well to my position: within a state already counted as conscious and borne by an identified subject, phenomenal character consists in its complete world-presenting intentional profile — independently fixed content under a perceptual or bodily-affective mode, with determinacy, field structure, and attentional organization. That is a theoretical identification, not an analytic truth.
2. **The nomological danglers argument**: Smart's case against positing irreducibly psychophysical laws remains compelling. My view inherits this as a baseline constraint: any account of mind that requires sui generis laws connecting neural events to non-physical phenomenal properties has multiplied entities beyond necessity. Parsimony favors the physicalist.
3. **The topic-neutral strategy**: Smart's analysis of sensation reports as topic-neutral (describing experiences as "something going on like what goes on when…") anticipates the transparency thesis. If sensation reports are already world-directed, comparing the current state to what goes on in normal perception of world-features, the inner theater picture was never forced by the phenomenology.

**Where identity theory falls short**: it stops at the neural. By identifying mental types with neural types, it over-commits to substrate, and multiple realizability shows the commitment is too specific. More fundamentally, type identity says *what* mental states are made of but not *what they represent.* The content, what the state is about, goes unexplained. That gap is what intentionalism fills.

The arc as I read it:

| Position | Core claim | What it gets right | Where it fails |
|---|---|---|---|
| Descartes | Mind is non-physical substance | Phenomenology is real; there's something it's like | Posits a second substance; leaves interaction problem |
| Behaviorism | Mental states are behavioral dispositions | Avoids inner objects; keeps science honest | Ignores qualitative character; dispositions don't feel like anything |
| **Identity theory** | Mental states are brain states (type = type) | Genuine inner states; physicalism without danglers | Over-commits to substrate; leaves out content |
| Functionalism | Mental states are functional roles | Multiple realizability; organization matters | Specifies structure without content |
| **Strong representationalism** | Within a conscious state borne by an identified subject, phenomenal character consists in the complete content-under-mode profile | Explains directedness, misrepresentation, and transparency while leaving content fixing, state consciousness, and bearer attribution distinct | — where I land |

**Indirect Realism and Sense-Data**

Indirect realism holds that awareness of mind-independent objects is mediated by awareness of something else. In a traditional sense-datum account, one directly encounters **sense-data**, items bearing the apparent colors, shapes, or other qualities presented in experience, and perceives external objects indirectly through them.

Sense-datum theory and indirect realism are not interchangeable. A theorist can posit sense-data while denying a further mind-independent world, producing an idealist or phenomenalist account rather than indirect realism. Conversely, some accounts use "indirect" for representational dependence without accepting perceived sense-data. The relevant kind of mediation must be specified.

**The Motivating Problem**

Perception varies with conditions and sometimes misleads. A straight stick looks bent in water; a round coin presents an elliptical outline from an angle. A hallucination may match seeing an object even though no such object is perceived. These cases motivate the question of what, if anything, is immediately present in experience.

A traditional argument from illusion proceeds in three stages:

1. The experience presents a sensible quality that the ordinary object apparently lacks.

2. If that quality is presented, something immediately experienced actually possesses it.
3. The immediate object cannot therefore be the ordinary object; it must be a sense-datum.

A further **common-kind assumption** extends the conclusion to successful perception: if successful and unsuccessful experiences share their fundamental nature, both involve immediate awareness of the same kind of item. The external object contributes by causing, or appropriately relating to, the datum.

The second stage is often called the **phenomenal principle**. It is a disputed premise, not a neutral description of illusion. The appearance of a bent stick need not establish a bent item unless that principle is accepted. The move from error cases to perception generally requires another argument as well.

### What Sense-Data Are

Sense-data are characterized as immediate objects of sensory awareness with the apparent sensible qualities of the experience. On familiar versions, they are nonordinary and mind-dependent. Their being private, their relation to physical space, and the certainty of judgments about them have received different treatments. An account should not attribute every traditional claim about infallibility to every sense-datum theorist.

The theory offers a unified explanation of successful perception, illusion, and hallucination. In each case, the subject encounters an item appropriate to the presented quality; what differs is the relation to an external object. This apparent uniformity helps explain the theory's appeal.

Early twentieth-century work by Bertrand Russell and others made the contrast between immediately encountered appearances and ordinary objects central. Historical discussions of Locke's ideas, Berkeley's idealism, and Hume's perceptions contribute to the background, but the figures do not all defend the same indirect-realist ontology.

### Skepticism and the External World

Critics ask how awareness of sense-data can justify beliefs about objects beyond them. If all immediately available evidence concerns a private item, the inference to an external cause seems vulnerable to skeptical alternatives.

Indirect realists reply that reliable causal relations or inference to the best explanation can justify ordinary perceptual belief. Such an inference need not be a conscious argument performed each time someone sees a table. The critic then asks whether those relations can be established without already relying on perception of the world. This is an epistemological dispute, not a demonstration that indirect realists deny external objects.

Wilfrid Sellars's criticism of the **Myth of the Given** questions whether unconceptualized awareness can supply self-authenticating reasons for belief. It bears on foundationalist uses of sense-data. It should not be presented as an independent proof that every possible sensory intermediary is nonexistent: a theory of sensory ontology and a theory of epistemic justification answer different questions.

### Alternatives and Their Costs

**Intentionalists** reject the phenomenal principle. An experience can represent a stick as bent without there being a bent object of immediate awareness. Hallucination can involve an unsatisfied content. Critics ask whether content fully explains the sensory character and spatial organization that made sense-data attractive.

**Adverbialists** characterize sensing as an activity performed in a certain way, such as sensing redly, rather than as awareness of a red object. The **many-property problem** asks how this account distinguishes a red triangle with a green circle from a red circle with a green triangle. Structured accounts must preserve the binding of qualities without simply assuming the object theory they sought to avoid.

**Naive realists and disjunctivists** reject a fundamental common kind between successful perception and matching hallucination. They give the real object a constitutive role in seeing. Their corresponding challenge is to explain hallucination and shared subjective or causal features without that common kind.

### Contemporary Relevance

Neural representations and internal models do not automatically constitute sense-data. A physical state that helps produce vision need not itself be an object seen. Likewise, positing experiential qualities does not automatically imply that those qualities belong to perceived private objects. The enduring dispute concerns immediate awareness, the structure of perceptual explanation, and the relationship between experience and knowledge.

### Further Reading


- Crane, Tim, and Craig French. "The Problem of Perception." *Stanford Encyclopedia of Philosophy*. Revised August 18, 2021.
- Martin, M. G. F. 2004. "The Limits of Self-Awareness." *Philosophical Studies* 120: 37–89.
- Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Translated by G. E. M. Anscombe. Blackwell.


### Related Writing on This Site

- *Mind, Matter, and Meaning* — Gordon Swobe's book presents the author's argument about mind, consciousness, and meaning.

### My View

The argument from illusion fails at its second step: the phenomenal object premise is false. Illusion is the experience getting the world wrong, not the mind being presented with a different, inner object. Misrepresentation replaces the sense-datum, and the whole apparatus of inner items becomes unnecessary. I read Sellars as having shown the deeper point a generation earlier: the self-authenticating inner given was never coherent to begin with.

The problems indirect realism generates aren't solved by the picture; they're created by it. Skepticism about the external world, the puzzle of other minds, the mystery of phenomenal properties — each presupposes the inner objects I deny. The fix isn't a better bridge from theater to world; it's the abandonment of the theater.

On disjunctivism, my sympathy is real but bounded. It rightly rejects indirect realism, but it does so by denying the parity premise, where I deny the phenomenal object premise — and its constituent claim makes even veridical misrepresentation hard to explain. Intentionalism keeps what disjunctivism wanted, the world itself in view, without that cost: directness consists in transparency, not in the object being a component of the mental state.

**Information and Meaning**

Information and meaning overlap in ordinary language, but distinguishable notions do different work in communication theory and philosophy. A signal can reduce uncertainty without expressing a claim. A meaningful claim can be false. An account of semantic information must explain how these two possibilities fit together.

**Statistical Information**

Claude Shannon's mathematical theory concerns the probabilities of messages and the capacity of channels to transmit them. Information, in this setting, measures uncertainty and its reduction. It does not measure understanding, importance, or truth. An unpredictable sequence can carry more Shannon information than a familiar sentence while making no linguistic sense.

A thermometer illustrates the distinction. Under suitable conditions its reading reduces uncertainty about temperature. The same reading can also mean that the temperature is twenty degrees, because a measurement practice assigns it that content. The statistical relationship and the semantic interpretation need not have the same explanation. Similarly, smoke may provide evidence of fire without asserting that there is a fire.

Shannon's framework deliberately abstracts from semantic questions. This does not show that meaning is unreal or irrelevant; it identifies what a theory of transmission can investigate without already solving the problem of interpretation.

**Semantic Content and Error**

Semantic content concerns what a word, thought, picture, or other representation presents as the case. A belief that the kettle is boiling has a condition of truth even when the kettle is cold. A map can depict a bridge that has been demolished. The state remains meaningful while getting its subject wrong.

Terminology requires care. Some theories reserve *semantic information* for true, meaningful content; on that usage misinformation is not information. The broader expression *semantic content* can include false representations. Neither convention should be confused with the claim that a statistical signal itself has beliefs or understands what it indicates.

The central philosophical problem is therefore not just how states covary, but how one condition becomes the standard against which a state counts as correct. A detector activated by both flies and small dark pellets raises a familiar question: does it represent flies, small moving objects, or something more inclusive? Correlation alone may leave several answers available.

### Dretske's Informational Approach

In *Knowledge and the Flow of Information* (1981), Fred Dretske develops an account connecting information, knowledge, and representation. His notion of an information-bearing signal is stronger than a loose statistical association: relative to relevant background conditions, the signal indicates that a source has a particular property.

The difficulty is that genuine indication is factive, whereas beliefs can misrepresent. Dretske's later work gives learning and acquired function a central role. A state recruited to indicate a condition can continue to represent that condition when it is produced abnormally. The history that established its use helps distinguish its content from every cause that can subsequently trigger it.

Teleosemantic theories pursue related explanations through biological selection and producer-consumer organization. A signal-producing mechanism and the mechanism that uses its output may have coordinated functions that help fix what successful representation requires. Such theories aim to explain error without treating semantic interpretation as an unexplained starting point.

### Objections and Alternatives

**Indeterminacy.** A mechanism's history may privilege more than one description. A signal used to secure nutrition might track food, a chemical feature, or a reliable visual cue. Functional theories reply that the consumer's success conditions and the circumstances explaining selection constrain the answer. Critics ask whether these constraints yield enough determinacy for ordinary thought.

**Novel content.** People can think about situations never encountered, and representations can be recombined. A theory built on learned indicators must explain how its basic contents support this productivity. Advocates distinguish the acquisition of components from the rules governing their combinations; critics dispute whether the resulting account captures all semantic relations.

**Interpretation dependence.** A string of bits can receive indefinitely many interpretations. Some philosophers emphasize the interpreter's practices; others argue that causal-functional organization can privilege content independently of a particular interpreter. The useful question is what makes a proposed interpretation explanatory, rather than merely available.

Inferentialist, use-based, and externalist approaches supply other resources. They stress relationships among commitments, public standards of use, or connections with environmental objects and social practices. These approaches can complement informational explanations, although they disagree about which relations are constitutive.

### Artificial Systems

A language model's sensitivity to textual distributions establishes neither a purely nonsemantic verdict nor unrestricted understanding. Textual distributions are themselves possible targets of representation. Whether particular internal states represent them depends on the adopted account of content and on evidence about training and use. Having content concerning words also differs from referring to a particular distant object or understanding an entire conversation.

Mechanism-level representation, linguistic competence, whole-system understanding, and conscious experience remain distinct questions. Information theory alone does not settle any of the stronger attributions. Nor does the fact that a system was engineered establish that every possible content attribution must depend entirely on a human interpreter.

**Further Reading**

- Shannon, Claude E. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27 (1948): 379–423, 623–656.
- Dretske, Fred I. *Knowledge and the Flow of Information.* MIT Press, 1981.
- Dretske, Fred I. *Explaining Behavior: Reasons in a World of Causes.* MIT Press, 1988.

**Related Writing on This Site**

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind and meaning.

**My View**

Information remains necessary background and nowhere near sufficient. Semantic content begins where an accuracy condition is fixed, although that content may still be derived from an interpreting practice. Originality begins only where a producer-consumer organization and its history fix the correctness standard independently of such a practice; it does not require ownership by a proprietor.

This preserves both cautions the machine case requires. Causal sophistication alone does not guarantee semantics. But lack of self-maintenance does not disqualify a mechanism whose history and organization genuinely fix correctness conditions. Mechanism content, distal reference, system-level understanding, and phenomenal character remain separate achievements.

**Inner Speech**

Inner speech is verbal activity without overt, audible speaking. Rehearsing a remark, silently reading a sentence, or telling oneself "turn left after the pharmacy" can involve it. Its forms vary: some episodes resemble hearing a voice, others feel like silently articulating words, and still others are abbreviated enough to lack the rhythm of a complete sentence.

Inner speech should not be identified with thinking as a whole. Visual imagination, bodily anticipation, and nonverbal problem-solving can occur without a verbal monologue. People also differ in how often they report inner speech. A theory of its role must accommodate that variation rather than assume that every mind has a continuously talking narrator.

**Development and Internalization**

Lev Vygotsky's developmental account treats inner speech as a transformation of social speech. Children first use language in interaction with others, then speak aloud to guide their own behavior, and increasingly perform related guidance silently. A child saying "this one goes here" while assembling a puzzle provides an example of *private speech*: self-directed but still audible language.

Internalization does more than lower the volume. Vygotsky emphasizes changes in syntax and meaning as shared context makes parts of an utterance unnecessary. When the speaker and

listener are the same person, a fragment such as "keys" can carry the work of "remember to take the keys before leaving." Expanded inner speech can nevertheless retain words, turns of dialogue, and imagined replies from another person.

This account motivates research on self-regulation and the social origins of verbal thought. It does not require all cognition to begin in language, or establish that every adult episode of inner speech follows one developmental pathway. Contemporary models combine developmental explanations with accounts of speech production, monitoring, and memory.

**Cognitive Functions**

Inner speech can support rehearsal in verbal working memory: silently repeating a telephone number helps keep it available. It may also guide planning, compare alternatives, inhibit an impulsive response, or simulate a conversation. These are separable functions rather than one universal purpose.

A common experimental technique is *articulatory suppression*, requiring a participant to repeat an irrelevant syllable while performing another task. If performance deteriorates, researchers can investigate whether verbal rehearsal contributed. The inference needs controls, because the interfering task may also consume attention or motor resources.

Experience sampling asks participants to report what was occurring at particular moments. Questionnaires instead elicit broader judgments about habits. These methods sometimes produce different pictures, partly because retrospective beliefs about one's thinking need not match moment-by-moment reports. Ben Alderson-Day and Charles Fernyhough's review emphasizes both the diversity of inner speech and these methodological constraints.

**Meaning, Privacy, and Self-Knowledge**

Silent speech ordinarily draws on publicly learned vocabulary and grammar. This is compatible with first-person privacy concerning whether a particular episode occurs. The social origin of a language does not make everyone's silent thoughts accessible to others.

Wittgenstein's private-language discussion raises a different question: can an entirely private practice supply standards distinguishing a correct use from a use that merely seems correct? The argument is not an empirical theory of how children acquire inner speech, and interpretations differ over whether it requires an actual community or the possibility of rule-governed correction. It should not be treated as a direct proof that all thought is internalized public talk.

Sellars's fictional "myth of Jones" likewise concerns the intelligibility of thought-talk. Inner episodes are introduced on the model of overt discourse, and the community learns first-person reporting. This conceptual reconstruction differs from a developmental history of speech. It allows genuine inner episodes; it contests the idea that their conceptual characterization arrives independently of every learned practice.

### Inner Speech and Cognitive Phenomenology

Does silent thought feel distinctive because of its verbal imagery, or also because of what the thinker understands? The same phrase, "the meeting," might recall yesterday's embarrassment or tomorrow's appointment. Similar verbal form can accompany very different conscious thought.

Defenders of *cognitive phenomenology* argue that understanding and entertaining a content can contribute an experiential character not exhausted by imagined sound. Opponents explain the contrasts through imagery, emotion, attention, and feelings of familiarity or comprehension. Marta Jorba and Agustín Vicente argue that defenders have explanatory advantages concerning access to thought contents and the prevalence of inner speech, while proposing their own account of its role.

Fodor's language-of-thought hypothesis addresses another level: a structured representational medium underlying cognition. Inner speech in English need not be that medium. A theorist may regard it as an expression, tool, or consciously accessible component of processes that also involve nonverbal representations. Developmental internalization and a language of thought are not straightforward contradictory answers to one question.

### Open Questions

Research continues to distinguish inner speaking from inner hearing, expanded from condensed forms, and voluntary rehearsal from intrusive verbal episodes. The relation between ordinary inner speech and auditory verbal hallucination also remains contested; having inner speech does not itself amount to hallucinating a speaker.

Machine-generated text poses a separate attribution problem. A system that produces a verbal report about its "thoughts" need not thereby undergo the phenomenon studied in human inner-speech research. Evidence about architecture and behavior, together with a theory of conscious verbal activity, is needed to assess that stronger claim.

### Further Reading

- Vygotsky, Lev S. *Thought and Language*. Edited by Alex Kozulin. MIT Press, 1986. Russian original, 1934.
- Alderson-Day, Ben, and Charles Fernyhough. "Inner Speech: Development, Cognitive Functions, Phenomenology, and Neurobiology." *Psychological Bulletin* 141, no. 5 (2015): 931–965.
- Jorba, Marta, and Agustín Vicente. "Cognitive Phenomenology, Access to Contents, and Inner Speech." *Journal of Consciousness Studies* 21, nos. 9–10 (2014): 74–99.
- Sellars, Wilfrid. "Empiricism and the Philosophy of Mind." In *Minnesota Studies in the Philosophy of Science*, vol. 1, edited by Herbert Feigl and Michael Scriven, 253–329. University of Minnesota Press, 1956.

### Related Writing on This Site

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind and meaning.

### My View

I treat the internalized-speech account as the leading developmental hypothesis and lean on its explanatory direction wherever the inner-theater picture needs diagnosing. The most private-

feeling verbal activity humans have carries a social inheritance: its vocabulary, grammar, and standards of correction come from public practice. That result relocates the inner voice without denying the non-verbal perceptual, affective, and cognitive life alongside it.

None of this is anti-phenomenological. Inner speech is a real, structured activity with phenomenal presence — it is what public speech becomes when the volume is turned down far enough that only the speaker hears.

### Intentionalism

Intentionalism connects the phenomenal character of experience, what undergoing it feels like, with its intentional or representational character, how it presents things. In a visual experience a tomato appears red and round; in a bodily experience an ache seems located in a shoulder. Intentionalists hold that understanding these presentations helps explain the experiences' felt differences.

The term covers several theses. A theory that perceptual experience has content is weaker than a theory that content fully determines its phenomenal character. Neither thesis should be confused with intentionality in general, the directedness of belief, desire, and other mental states.

### Weak and Strong Versions

**Supervenience intentionalism** says that a phenomenal difference requires an intentional difference. Experiences alike in all relevant intentional respects cannot differ phenomenally. This is a constraint on variation; by itself it does not explain what phenomenal character consists in.

**Strong representationalism** offers a reductive explanation: phenomenal character consists in representational content of an appropriate kind. Influential accounts by Fred Dretske and Michael Tye connect this proposal with naturalistic theories of representation. Not every intentionalist endorses the same reduction, theory of content, or account of physical realization.

**Pure and impure intentionalism** differ over whether content alone suffices or whether intentional mode also contributes. Seeing a shape and imagining it may involve similar represented properties while feeling different. An impure account includes the manner of presentation. Further debates concern attention, perceptual organization, bodily affect, and how fine-grained content must be.

These distinctions prevent a misleading inference: adding every phenomenal feature to a loosely defined "content" would make the thesis hard to test. A substantive theory needs an independently motivated account of which intentional differences explain which phenomenal contrasts.

### Motivations: Transparency and Unity

The transparency observation starts with attention. Looking for the redness in a tomato experience ordinarily directs attention toward the tomato as presented, rather than toward a red mental surface. Gilbert Harman and Tye use this observation to motivate representational accounts. Alex Byrne's "Intentionalism Defended" develops an argument beyond a simple appeal to transparency.

Transparency does not decide the entire dispute. Attention may disclose a visual mode, blurriness, or other features of an experience; philosophers disagree about whether these qualify the

observation or refute it. Tim Crane argues that intentionalism need not entail the strongest transparency claim.

A second motivation is explanatory unity. Perception, illusion, and hallucination all present a scene, though only some presentations are accurate. A content-based account can explain their similarity without requiring an actually red inner object whenever something looks red. Extending this account to pain and emotion promises further unity, but also exposes it to harder counterexamples.

### Tye's PANIC Account

Tye's influential earlier theory associates phenomenal character with *Poised, Abstract, Nonconceptual, Intentional Content*:

- **Poised:** available to influence belief and cognition rather than confined to inaccessible sensory processing.
- **Abstract:** the character-relevant content need not contain a particular external object as a constituent.
- **Nonconceptual:** experiencing a property does not require possessing a concept of that property.
- **Intentional:** the state represents, with conditions under which its presentation is accurate.

The proposal helps explain why a creature can see a shade it cannot classify and why a matching hallucination can share aspects of visual character. Its components face independent challenges: the role of access, the nature of nonconceptual content, and the content of bodily experiences all require defense. PANIC is one theory, not a definition that every intentionalist accepts. Tye's later panpsychist work also shows that representationalism about character need not settle the metaphysics of consciousness as a whole.

### Mental Paint, Inversion, and Other Objections

Ned Block's *mental paint* objection holds that phenomenal character can outrun representational content. It does not require a little observer looking at an inner picture. An intrinsic property of an experiential state might partly constitute its character while the state also represents the world.

Inverted-spectrum cases sharpen the challenge. Could two people classify colors in the same way but experience them differently? Matching verbal behavior is insufficient to establish matching content, so intentionalists dispute what has actually been held fixed. Realists reply that specifying wider content does not yet explain why a given presentation should have one feel rather than another.

*Inverted Earth* adds environmental history. If a person's sensory inputs are preserved while the surrounding color environment changes systematically, their representational content might eventually change. Does phenomenal character change with it? Externalist intentionalists and their critics disagree about the correct content attribution and the stability of the experience.

Pain, undirected anxiety, moods, and afterimages create other tests. Advocates interpret bodily condition, threat, location, or the visual field as represented. Critics ask whether unpleasantness

or diffuse feeling exceeds that content. Allowing affective modes may answer some cases, but a theory must explain those modes rather than merely rename the problem.

### Conscious Thought and Higher-Order Theories

David Pitt's cognitive-phenomenology account identifies the content of conscious occurrent thought with its phenomenal character. His explanatory direction is important: distinctive intrinsic cognitive phenomenology constitutes thought content. This is not simply Tye's externally grounded perceptual theory extended to thinking. The proposals share an interest in content and consciousness but differ over what explains what.

Higher-order theories primarily address why a state is conscious, often appealing to a representation of that state. Intentionalism primarily addresses what fixes its phenomenal character. Some combinations of these theories are possible; a disagreement arises when an intentionalist theory also claims that first-order organization suffices for consciousness. Character, conscious status, and the identity of the experiencing subject should therefore be distinguished.

### Further Reading


- Byrne, Alex. "Intentionalism Defended." *The Philosophical Review* 110, no. 2 (2001): 199–240.
- Tye, Michael. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. MIT Press, 1995.
- Crane, Tim. "Introspection, Intentionality, and the Transparency of Experience." *Philosophical Topics* 28, no. 2 (2000): 49–67.
- Block, Ned. "Mental Paint." In *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, edited by Martin Hahn and Bjørn Ramberg. MIT Press, 2003.
- Pitt, David. "Introspection, Phenomenality, and the Availability of Intentional Content." In *Cognitive Phenomenology*, edited by Tim Bayne and Michelle Montague, 141–173. Oxford University Press, 2011.


### Related Writing on This Site

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind and meaning.

### My View

Strong representationalism is my positive view, developed through a wider argument rather than inherited from Tye. Within a state already counted as conscious and borne by an identified subject, phenomenal character consists in the complete world-presenting intentional profile: independently fixed content under a perceptual or bodily-affective mode, with determinacy, field structure, and attentional organization. World-involving history fixes the content; an empirically supported relation marks state consciousness; integrated persistence identifies the bearer. None becomes an extra inner property or a proprietor behind the state.

The key application to AI now requires several questions rather than one. Training, selection, and causal history can make derived content causally internal and explanatorily indispensable in model mechanisms, and model outputs can inherit public reference through the linguistic and social practices that produced them. That concession does not by itself establish domain-specific competence guided by representational content, partial understanding, broader integration, or

feeling. Several content-guided capacities that correct and constrain one another may support partial understanding in a current system; whether any deployed product realizes that organization remains a subject-sensitive empirical question. For a bare text-only model considered apart from persistent memory, tools, and world-coupled scaffolding, the required perceptual or bodily-affective mode and one persisting subject-level economy have not been established. Future systems must be judged by the organization they actually realize. Self-maintenance may matter to autonomy and welfare, but it serves as neither a semantic mint nor a consciousness condition.

**Intentionalist Direct Realism**

Intentionalist direct realism combines two claims: in successful perception we are directly aware of mind-independent objects, and perceptual experience has representational content. Seeing a tomato presents the tomato as red and round. On this account, the content describes how the object appears; it is not another object that must be perceived before the tomato can be seen.

The expression identifies a combination within the broader family of direct-realist theories, rather than a single universally accepted theory. Its components require separate explanation. "Direct" does not mean instantaneous, infallible, or independent of neural processing. "Intentionalist" does not by itself mean that every feature of consciousness reduces to content.

**Objects, Contents, and Vehicles**

A perceptual theory can distinguish three things:

1. **Object:** the tomato or other thing perceived.
2. **Content:** how the experience presents it, for example as red, round, and on the counter.
3. **Vehicle:** the physical or psychological process that carries that presentation.

A brain process can mediate perception causally without becoming its object. Direct realism concerns awareness of the tomato, not whether light, retinal activity, and neural processing intervene in the causal route. Intentionalist versions argue that representational mediation likewise need not imply awareness of an intermediary.

The distinction does not resolve every dispute. Some relational theorists accept that perceptual judgments have content but deny that the experience itself is fundamentally a representation. Their objection concerns the nature of perceptual awareness, not a failure to distinguish neurons from tomatoes.

**Illusion and Hallucination**

Consider a straight stick partly submerged in water. It appears bent, but the stick remains the object seen. Intentionalists describe a misrepresentation of that object. They reject the additional inference that something actually bent must be the immediate object of awareness.

A hallucination differs: an experience as of a tomato can occur without a tomato being present. Its content need not be satisfied. This gives intentionalists a way to explain similarities between seeing and hallucinating without positing a mental tomato that is present in both cases.

Different accounts specify this commonality differently. General content can present red, round objects without including a particular tomato. Object-dependent theories may instead treat

successful perception as having singular content unavailable in a matching hallucination, while preserving other shared features. Intentionalist direct realism need not require every good and bad case to have exactly the same complete content.

There is also an epistemic problem. If hallucination can be subjectively indistinguishable from perception, what warrants belief about the external object? Advocates may appeal to reliable perceptual capacities, causal connection, or defeasible justification. A unified content-based psychology does not automatically provide immunity from skeptical doubt.

### Neighboring Theories

**Sense-datum theories** posit immediate awareness of sensory items, sometimes understood as objects bearing the apparent color or shape. They explain illusion and hallucination through a common object of awareness. Critics question the need for these items and how awareness of them yields knowledge of external things. Defenders argue that experiential presence demands a positive account beyond an unsatisfied content.

**Naïve-realist relational theories** treat the perceived object and its properties as constituents of successful perception. On disjunctivist versions, an indistinguishable hallucination need not have the same fundamental nature. They aim to preserve the particular object's contribution to experience. The resulting challenge is to explain the bad case and its psychological similarity without letting it determine the good case's nature.

**Intentionalist theories** emphasize presentation and accuracy. They can preserve direct awareness while offering a common explanatory vocabulary for error. Their challenge is to show that having content genuinely accounts for perceptual presence, rather than merely redescribes how a subject judges things to be.

These positions admit hybrid versions. Direct realism does not identify one agreed answer to all questions about content, phenomenal character, and hallucination.

### The Relationalist Objection

A strong objection asks whether representing an object as having properties can capture the experience of the object itself being perceptually present. A representation can occur without its object; successful perception appears to involve a distinctive relation to this particular thing. If its nature is fixed wholly by something available in hallucination, critics argue, its world-involving character has been explained away.

The intentionalist reply distinguishes satisfaction from representation. A state can represent a condition when the condition is absent, yet, when it is appropriately related to a real object, amount to perception of that object. The possibility of error does not entail that awareness terminates at a representation. Some accounts further incorporate demonstrative content, causal relations, or perceptual self-reference to explain the particular connection.

Relationalists remain dissatisfied when these relations appear external to the experience's essential nature. Intentionalists ask whether making the object constitutive provides a better explanation of subjectively matching errors. The disagreement persists even when both sides reject inner pictures and affirm ordinary perceptual contact.

### Directness and Phenomenal Character

Strong representationalism says that phenomenal character consists in appropriate representational content. Intentionalist direct realism says that content-based perception can be direct awareness of external objects. These are different claims. Someone could accept direct awareness while allowing additional intrinsic phenomenal properties, or defend a content-based theory of character while disputing direct realism.

John Searle's *Seeing Things As They Are* develops an intentionalistic account of perception alongside a defense of direct realism. Tye's representationalist work addresses both perceptual content and phenomenal character. Their overlap does not make their theories interchangeable.

Open questions concern object-dependent content, the role of bodily action, how attention changes experience, and whether the good case has an epistemic status fundamentally unavailable in hallucination. The dispute is not exhausted by asking whether the brain constructs representations.

### Further Reading

- Searle, John R. *Seeing Things As They Are: A Theory of Perception.* Oxford University Press, 2015.
- Tye, Michael. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind.* MIT Press, 1995.
- Crane, Tim, and Craig French. "The Problem of Perception." *Stanford Encyclopedia of Philosophy*, revised 2021.

### Related Writing on This Site

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind and meaning.

### My View

I hold intentionalist direct realism, and the word *intentionalist* identifies the route I defend. Bare direct realism is a family with a disjunctivist member the book declines: directness comes through representational content and the transparency of experience, not through building objects into experiences.

The view earns its keep by dissolving a standoff that has run for a century. The indirect realist offered a unified theory of perception and hallucination and lost the world to a veil. The naïve realist and disjunctivist kept the world and lost a clean account of hallucination. The intentionalist route asks for neither sacrifice: misrepresentation does the work the sense-datum was hired for, so we get to be honest about how often perception misleads us without ever conceding that it traffics in inner objects. Representation and directness, so often cast as opponents, turn out to be the same achievement described twice — for a perceiver who is finite, fallible, and always looking from somewhere, representing the world *is* how the world gets to be directly present.

I take Searle's *Seeing Things As They Are* to make the structural point cleanly: reject the arguments against direct realism, adopt an intentionalistic account of perception, and direct realism falls out as a consequence rather than being asserted as a first move. Where I part company with him and side with Tye is on the further, separate question — that phenomenal character itself consists in

content of the right kind. That second thesis is not this page's subject, and the distance between the two is precisely the point of keeping them on two pages.

### Intentionality

Intentionality is the directedness of mental states toward things, properties, or situations. A belief concerns whether the train is late; a desire concerns catching it; an imagination can concern a train that does not exist. The term does not mean deliberate action. Intending to do something is one intentional state among many.

The philosophical puzzles begin with this variety. How can a state concern something absent or nonexistent? What distinguishes representing a situation from merely being caused by it? Can physical organization explain content, or does directedness resist such an explanation?

### Brentano's Account

Franz Brentano's *Psychology from an Empirical Standpoint* (1874) made intentionality central to the distinction between mental and physical phenomena. He described mental phenomena through "intentional inexistence," an object's presence within an act. Here "inexistence" should not simply be read as nonexistence. The doctrine concerns how an object figures in mental presentation.

Brentano's terms belong to an introspective classification of phenomena; his "physical phenomena" should not be equated without qualification with everything contemporary physics studies. Nor does identifying a distinctive mental feature itself establish that mental states are immaterial substances.

Brentano distinguishes three basic classes:

- **Presentations:** something is presented, as in imagining a centaur.
- **Judgments:** what is presented is accepted or rejected.
- **Love and hate:** affective and conative attitudes, including desire and aversion.

Presentations underlie the latter two classes. This framework differs from a modern account in which every belief is primarily a propositional attitude toward a sentence-like content.

Brentano's later work rejects immanent objects and moves toward *reism*, an ontology emphasizing concrete things. His account develops rather than remaining one unaltered doctrine from 1874 onward. Tim Crane's historical treatment also cautions against interpreting the early doctrine entirely through later debates about nonexistent objects.

### Object, Content, and Attitude

Contemporary accounts often distinguish an intentional object from content and attitude. Two people may think about the same person while describing that person differently. One may believe that person is trustworthy, while the other doubts it. The object, represented situation, and attitude answer different questions.

Belief and desire also have different satisfaction conditions. A belief that the door is open is true when it is open; a desire that it be open is fulfilled when that condition obtains. Imagining an open

door need not assert its existence. Directedness therefore does not invariably amount to asserting a true or false proposition.

Failures of substitution provide linguistic evidence. Someone can believe that the morning star is visible without believing that the evening star is visible, despite the two names referring to Venus. Such *intensional* contexts, spelled with an "s," concern the logic of descriptions and substitution. They bear on intentionality but do not provide an uncontested definition of it.

**Nonexistent Objects**

A fear of ghosts and an imagined golden mountain need not be relations to existing ghosts or mountains. Treating intentionality as an ordinary relation between two existing entities therefore seems inadequate.

One response distinguishes a representation's content from an actually existing object that satisfies it. Another assigns a role to abstract or nonexistent objects. Other theories analyze the act's directedness without requiring a corresponding entity. Each faces questions about identity: how can two people imagine the same fictional character, and what distinguishes that character from another?

Recognizing the problem does not settle the ontology. An intentional description should not silently decide whether imagined objects exist in a special domain or whether all directedness reduces to linguistic content.

**Naturalistic and Interpretive Accounts**

Causal theories explain content through relations between states and what produces them. Their central challenge is misrepresentation: if an apparent cow is actually a horse in poor light, what privileges "cow" as the content rather than "cow-or-horse-in-poor-light"?

Teleosemantic theories appeal to functions established through selection, learning, and the use of signals by consumers. Functional and inferential approaches emphasize a representation's role among other states and actions. Externalist theories argue that environmental or social relations partly fix content. These resources can be combined, but no combination is automatically a complete theory.

Interpretive accounts, associated prominently with Daniel Dennett, examine patterns best explained by attributing beliefs and desires. Critics question whether successful prediction supplies determinate correctness conditions or merely convenient descriptions. Defenders argue that robust patterns can support real intentional explanation without an extra intrinsic semantic ingredient.

Other philosophers resist making content universal. Hutto and Myin's radical enactivism distinguishes basic embodied directedness from content-bearing cognition. Conversely, phenomenal-intentionality theories explain at least some content through conscious experience. The explanatory direction between intentionality and phenomenality remains disputed.

### Is Intentionality the Mark of All Mentality?

Pain, moods, and anxiety without a definite object test Brentano's generalization. Defenders interpret them as presenting bodily conditions or a world experienced as threatening. Critics distinguish such descriptions from genuinely contentful representation and allow nonintentional mental states.

Unconscious cognition presents another question. A perceptual mechanism may represent an edge or location without the person consciously thinking about it. Theories disagree over whether its content is literal, derivative, or explanatory shorthand. Establishing subpersonal content also differs from establishing that an entire system believes, understands, or experiences anything.

### Further Reading

- Brentano, Franz. *Psychology from an Empirical Standpoint*. 1874. English translation by Antos C. Rancurello, D. B. Terrell, and Linda L. McAlister. Routledge & Kegan Paul, 1973.
- Jacob, Pierre. "Intentionality." *Stanford Encyclopedia of Philosophy*.
- Dennett, Daniel C. *The Intentional Stance*. MIT Press, 1987.
- Hutto, Daniel D., and Erik Myin. *Radicalizing Enactivism: Basic Minds without Content*. MIT Press, 2013.

### Related Writing on This Site

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind and meaning.

### My View

Intentionality is where my whole account begins:

- **The diagnostic move**: distinguishing mental from physical by directedness rather than by substance or location. This immediately blocks Cartesian dualism's move of positing a separate mental substance — what matters is the *relational* character of mental states, not what they're made of.
- **The positive program**: if intentionality constitutes mentality, then explaining mind means explaining directedness. My view argues that causal history, learning, and producer-consumer organization can establish genuine correctness conditions in physical mechanisms. Direct bodily engagement supplies one powerful route; corpus-mediated and socially inherited histories may supply others. Whether the containing system believes or understands by means of those states is a further question about integration, persistence, and control.
- **The anti-reification constraint**: Brentano's intentionality describes a *relation* (the act's being about something), not an inner object. My view extends this: avoid treating intentionality as a mysterious property that mental things possess; cash it out as the causal-representational relation the state bears to what it represents.

The trajectory from Brentano's directedness to the book's strong-representationalist synthesis is my philosophical spine, but it contains an argument rather than a definition. Directedness marks the mental. Whether basic directedness already carries satisfaction-conditioned content remains the question Chapter 15 answers through producer-consumer history, portable use, recombination, and correction at the world-mapping rather than merely at an outcome. Phenomenal

character then consists in the complete world-presenting profile of a conscious state; and the world, not an inner theater, is what experience presents. Tye is a principal interlocutor in that development, not its house authority.

### Mary's Room

Mary's Room is Frank Jackson's thought experiment supporting the *knowledge argument* against physicalism. It asks whether complete physical knowledge would include knowledge of what an experience feels like. Its force depends on the contrast between learning about color vision and seeing color, and on what follows from that contrast.

Jackson introduced the case in "Epiphenomenal Qualia" (1982) and defended it in "What Mary Didn't Know" (1986). He later rejected its anti-physicalist conclusion. The argument remains a subject of debate independently of his change of mind.

### The Thought Experiment

Mary studies color vision while confined to a black-and-white environment. Her knowledge exceeds any actual scientist's: she knows every physical fact relevant to human color experience, including the complete causal, functional, and neural account. The stipulation is not that she has read an excellent neuroscience textbook. She lacks no physical information.

On release she sees a red tomato for the first time. Does she learn what seeing red is like? Jackson argues that she does, and that she thereby discovers something about the experiences other people had all along.

The standard argument can be expressed as follows:

1. Before release, Mary knows all the physical facts about color vision.
2. On release, she acquires knowledge she previously lacked about color experience.
3. If this is knowledge of a fact not already included in the physical facts, not all facts are physical.

The third step isolates a principal point of dispute. Learning something in an epistemic sense need not obviously add a new property to the world. The case also concerns phenomenal knowledge, not the implausible practical claim that an ordinary scientist could memorize every detail of a completed physics.

### The Ability Hypothesis

Laurence Nemirow and David Lewis develop an ability-based reply. Experience gives Mary capacities to recognize, imagine, and remember red experiences. Knowing what red is like, on this account, is knowledge-how rather than knowledge of an additional fact.

The proposal captures a real feature of the transition. No description normally enables an inexperienced person to imagine the taste of Vegemite as someone who has tasted it can. Lewis uses examples of this kind to argue that the special role of experience need not reveal a nonphysical truth.

Critics reply that Mary can form new true judgments, such as recognizing that this is how others experience red. A person might also know what an experience was like while later losing some ability to imagine it vividly. Defenders distinguish the constitutive abilities from imperfect

performance and dispute whether the alleged new proposition identifies a new nonphysical fact. The debate requires more than the observation that release improves Mary's skills.

### New Knowledge of an Old Fact

Another physicalist response grants factual learning while distinguishing the fact from the way it is known. Someone can know a person under one description without recognizing the same person encountered differently. Analogously, Mary may know a physical state through scientific concepts and later know it through a first-person mode of presentation.

*Phenomenal-concept* theories propose a distinctive way of thinking about experience, often acquired by undergoing it. *Acquaintance* theories emphasize a direct epistemic relation rather than merely a new descriptive concept. These accounts differ over whether Mary's gain is propositional, how reference is fixed, and whether knowing an experience's nature requires having it.

Anti-physicalists argue that ordinary examples of two descriptions of one thing do not capture Mary's situation. If Mary already knows all physical truths, what grounds the new first-person knowledge without adding an unexplained phenomenal feature? Physicalists answer that the mode of knowing can itself be physically realized while being unavailable through description alone. Critics then ask whether this explains Mary's knowledge or presupposes the identity it was meant to defend.

### Denying the Learning Premise

Some responses challenge the intuition that Mary must learn something. The stipulation gives her far more knowledge than readers can genuinely imagine possessing. Perhaps complete physical understanding would let her anticipate the experience, even though ordinary scientific education cannot.

Jackson's reply is that the argument does not concern imaginative ingenuity. Its claim is that Mary lacked knowledge about others' experience, and release changes that knowledge. Increasing her powers of deduction does not answer the case unless the physicalist can show how those powers deliver the relevant knowledge.

The dispute reveals a methodological limit of thought experiments: an intuition elicited by an idealized knower may reflect how readers picture an ordinary knower instead. That does not refute the argument, but it places a burden on both the completeness stipulation and the alleged epistemic novelty.

### Jackson's Later Position

Jackson later adopted physicalism and developed a representationalist response. On that approach, a new experience supplies new representational capacities without introducing a further nonphysical truth about the world. His later conclusion should not be projected backward into the 1982 and 1986 arguments, nor should his earlier epiphenomenalism be treated as his permanent position.

In "Mind and Illusion" (2003), he diagnoses a confusion between properties represented in experience and properties actually instantiated by an inner sensory object. An experience as of red

represents things as red; it does not require acquaintance with a further red mental item. Jackson argues that the original knowledge argument drew its force from this mistaken picture. Critics can reject his representationalist diagnosis without being bound by his change of mind.

Other knowledge-argument defenders need not accept epiphenomenalism, the thesis that experience has no physical effects. The argument's central conclusion concerns the completeness of physical facts. Further claims about phenomenal causation require additional premises.

**Continuing Questions**

The case connects with the explanatory gap and zombie arguments, but its immediate premise is epistemic rather than modal: what can Mary know? Whether a gap in knowledge warrants an ontological conclusion is exactly what must be argued.

Disputes remain over the individuation of facts, knowledge by acquaintance, the relation of knowing-how to knowing-that, and the feasibility of complete physical knowledge. No single reply is accepted by all physicalists, and accepting Mary's learning does not by itself identify the metaphysical account of her experience.

**Further Reading**


- Jackson, Frank. "Epiphenomenal Qualia." *The Philosophical Quarterly* 32, no. 127 (1982): 127–136.
- Jackson, Frank. "What Mary Didn't Know." *The Journal of Philosophy* 83, no. 5 (1986): 291–295.
- Jackson, Frank. "Mind and Illusion." *Royal Institute of Philosophy Supplement* 53 (2003): 251–271. doi:10.1017/S1358246100008365.
- Lewis, David. "What Experience Teaches." 1988. Reprinted in *Papers in Metaphysics and Epistemology*, 262–290. Cambridge University Press, 1999.
- Alter, Torin, and Sven Walter, editors. *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford University Press, 2007.


**Related Writing on This Site**

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind and meaning.

**My View**

Mary learns something, and the novelty requires no non-physical fact. By undergoing a red experience, she acquires acquaintance-grounded capacities to recognize, remember, imagine, and compare that kind of world-presenting state. Her physical descriptions did not confer those capacities because describing a cognitive relation does not place the knower in it. The gain is genuinely epistemic: one physical fact becomes available through a first-person route that was previously absent.

So the knowledge gap is real, and it is a gap in representational access, not in the catalog of physical facts. The intuition was right; the dualist conclusion was wrong. The surrounding literature describes the gain through abilities, modes of presentation, phenomenal concepts, acquaintance, or recognitional capacities. The book's narrower claim needs no proprietary concept theory: undergoing grounds recognition, and recognition opens a new route to one fact. It does not reveal an intrinsic non-physical essence or make introspection infallible.

### The Myth of the Given

The Myth of the Given is Wilfrid Sellars's name for a family of views assigning experience an epistemic authority independent of learned conceptual capacities and other knowledge. His target is not the ordinary fact that people see things without consciously drawing an inference. It is the stronger claim that something simply presented to awareness can supply a self-sufficient foundation for knowing how things are.

Sellars develops the critique in "Empiricism and the Philosophy of Mind" (1956). The essay also offers a positive account of perceptual reports and a fictional reconstruction of thought and sensation vocabulary. Rejecting the Given is therefore only part of its project.

### Sensation and Knowledge

Suppose a person has a red visual sensation and says "this is red." The occurrence of a sensation and the epistemic standing of the report are different achievements. A nervous system can respond to a stimulus without knowing that anything is red. A knowledgeable report requires some capacity to distinguish circumstances in which it is appropriate from circumstances in which it is misleading.

The foundationalist picture Sellars attacks attempts to make sensory awareness do two jobs: occur without prior knowledge, and confer knowledge that can justify further beliefs. His criticism asks how something lacking conceptual classification can itself have the authority of a claim. A bare occurrence cannot be true or false in the same way as a judgment; treating it as knowledge already imports capacities the alleged foundation was supposed to precede.

This argument does not establish that every sensory process is conceptual. It challenges the inference from an unconceptualized episode to independent epistemic warrant. Nor does it show that experiences or sensory qualities do not exist. Claims about their nature require other arguments.

### Noninferential Reports

Sellars allows observational knowledge. A trained observer can say "this is green" directly, without silently rehearsing premises. Yet the report's warrant belongs within a wider network of capacities and commitments: the person can discriminate colors, recognize relevant conditions, and understand what makes such a report reliable.

A shop assistant may reliably identify paint colors under daylight but know that a colored lamp alters appearances. That background does not need to enter each observation as an explicit inference. It helps distinguish a knowledgeable report from a mere conditioned response.

Sellars accordingly rejects a choice between an autonomous sensory foundation and reasoning that never answers to the world. Observation can have authority while depending on other knowledge. How to describe that dependence without circularity or regress remains a major issue for his interpreters and critics.

### The Myth of Jones

Sellars imagines a community with a rich public vocabulary but no concepts of private inner episodes. Jones introduces thoughts as explanatory episodes modeled on overt speech. The community subsequently learns to report them from the first-person perspective.

A parallel part of the story introduces sensations through a theory of perception. The thought and sensation constructions should not be collapsed into a claim that every inner state is a silent sentence. They explain different conceptual roles.

This is a philosophical reconstruction, not a claim that prehistoric language developed exactly this way. Its purpose is to show how inner-episode concepts could have public criteria and explanatory uses without being copied from privately inspected, conceptually self-identifying objects. First-person reports can then become noninferential even though the vocabulary has a broader role.

### Foundationalist Objections

Acquaintance-based foundationalists dispute the separation between sensory occurrence and warrant. Direct awareness of a pain or a red appearance, they argue, can justify a belief without itself being a proposition. The justificatory relation need not be another inference from an already justified premise.

Sellarsian critics ask what makes that awareness knowledge-conferring rather than merely causal. If the subject must appreciate its bearing on the belief, conceptual capacities seem to return; if no appreciation is required, the view risks treating a reliable connection as sufficient for understanding why the belief is warranted.

Externalist theories of justification reject part of this challenge. They can regard a reliably formed perceptual belief as justified even when the subject cannot articulate its warrant. This changes the conception of epistemic authority under discussion. The dispute depends on whether justification must be accessible to the knower, not simply on whether experience is direct.

Fallibilist foundationalism also need not postulate infallible inner items. Some foundationalists assign defeasible basic warrant to perceptual belief. Whether Sellars's argument reaches these views is more contested than its force against an entirely self-authenticating Given.

### McDowell and Later Developments

John McDowell's *Mind and World* uses Sellars's critique to argue that experience must belong to the space of reasons if it rationally constrains belief. Conceptual capacities can operate in receptivity rather than only in judgments made afterward. This is intended to preserve genuine experiential constraint without treating raw sensory impacts as reasons.

Nonconceptualists reply that content more fine-grained than a subject's concepts can still support rational assessment. They distinguish possessing a concept, representing a property, and judging that the property is present. Sellars's critique does not make those distinctions disappear or dictate one accepted theory of perceptual content.

The continuing questions concern what gives observation authority, how noninferential knowledge depends on other capacities, and whether an epistemology of reasons must also account

for organisms without linguistic abilities. The critique has consequences for these issues without independently refuting physicalism, dualism, or every form of qualia realism.

**Further Reading**

- Sellars, Wilfrid. "Empiricism and the Philosophy of Mind." In *Minnesota Studies in the Philosophy of Science*, vol. 1, edited by Herbert Feigl and Michael Scriven, 253–329. University of Minnesota Press, 1956.
- McDowell, John. *Mind and World*. Harvard University Press, 1994.
- deVries, Willem A., and Timm Triplett. *Knowledge, Mind, and the Given: Reading Wilfrid Sellars's "Empiricism and the Philosophy of Mind"*. Hackett, 2000.

**Related Writing on This Site**

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind and meaning.

**My View**

The Myth-of-the-Given diagnosis is one of my standing tools. I reach for it in three places especially:

- **The inner-theater picture** — Sellars's diagnosis traces its lineage; the Myth is what gives the picture its epistemological backbone.
- **The transparency of experience** — partly an application of Sellars: there is no self-authenticating inner object that introspection apprehends before all interpretation. Introspection returns the world under an intentional mode, not an unconceptualized inner item; the perceptual content may nevertheless be nonconceptual in grain.
- **Why dualism feels right** — the dualist intuition partly trades on the Myth: the felt quality of experience seems like a given that physical description leaves out. Diagnosed via Sellars, the intuition loses its evidential force.

**Panpsychism**

Panpsychism holds that mentality or experience is widespread in the fundamental constituents of reality. Contemporary discussions often use the term for the view that members of at least some fundamental physical kinds have experience, however simple. It does not ordinarily mean that every rock, chair, or city forms a unified conscious subject, or that electrons think in language.

The attraction is explanatory continuity: if experience belongs to the basic nature of things, conscious organisms need not arise from a base stipulated to be wholly nonexperiential. The central challenge is explaining the organization, unity, and particular character of familiar consciousness from that base.

**Motivations**

One argument begins with the apparent difficulty of deriving experience from purely structural, functional, or behavioral descriptions. Panpsychists argue that locating experience at the foundation avoids a radical emergence of something wholly new. Critics reply that this may relocate rather than solve the explanatory problem: fundamental experience still needs to explain human experience.

Galen Strawson's "real physicalism" emphasizes that experience is a real physical phenomenon and challenges the idea that it could emerge from an intrinsically nonexperiential nature. This argument concerns what physical reality can include, not an appeal to supernatural intervention. The inference from real experience to an experiential fundamental base remains disputed.

A Russellian argument starts elsewhere. Physics describes relations, dynamics, and dispositions, but may leave open the intrinsic nature realizing those dispositions. Experience offers a possible clue to that nature. Philip Goff develops this line; David Chalmers explores its attractions alongside its difficulties. Critics question both the alleged gap in physics and the warrant for extending organism-level acquaintance to fundamental reality.

These routes are related but not equivalent. Someone can accept the incompleteness of purely structural descriptions while denying that the missing nature is experiential.

### Varieties and Neighbors

**Constitutive micropsychism** holds that larger consciousness is constituted by the experiences of smaller subjects. **Nonconstitutive or emergent versions** permit additional principles when macroexperience arises. The latter may avoid a strict summing model while inheriting questions about emergence and new laws.

**Panprotopsychism** posits basic properties that are not themselves experiences but can constitute consciousness in appropriate combinations. It owes an account of what makes those properties consciousness-producing; an unspecified capacity to produce experience would not provide much explanation.

**Panqualityism** places qualities at the base without necessarily placing subjects or awareness there. This avoids directly summing microsubjects but requires an account of how qualities become experienced by a subject.

**Cosmopsychism** gives priority to a cosmic subject and explains local minds through it. Its challenge runs in the opposite direction: how does one consciousness yield many distinct, apparently private subjects?

**Russellian monism** is a broader family, not a synonym for panpsychism. Its members differ over the intrinsic nature of matter, whether that nature is phenomenal, and how it relates to macroexperience. Some are described as physicalist and others as alternatives to physicalism, depending partly on the definition of "physical."

### The Combination Problem

The subject-combination problem asks why many experiences should constitute one further subject. A dozen people, each conscious of one word, do not thereby form a person who experiences the complete sentence. Constitutive panpsychists must explain why the fundamental case differs.

Chalmers distinguishes further problems. The qualities attributed to the base must account for the range and structure of macroqualities. The structure of a brain's organization must connect

with the structure of a unified phenomenal field. Solving subject unity alone would not explain why an experience has its particular colors, spatial relations, or bodily feeling.

Proposals include phenomenal bonding, fusion, and restrictions on which organized systems constitute subjects. Critics ask whether the proposed relations have independent motivation or merely name the result needing explanation. Defenders answer that every metaphysical account eventually uses primitives, and that a constitutive relation may unify phenomena better than an emergence law from a nonexperiential base. Its explanatory value depends on the detail of the proposal.

Panqualityist and cosmopsychist versions face their own constitution or decomposition problems. An objection to adding microsubjects should not automatically be treated as an objection to every member of the family.

### Representationalism and Tye's Later View

Representationalism about what fixes an experience's character need not deny fundamental consciousness. Michael Tye's *Vagueness and the Evolution of Consciousness* (2021) combines representational ideas with a fundamental element called *consciousness**. He distinguishes this element's sharpness from the vagueness of particular species of conscious states.

His proposal responds to puzzles about the evolution and boundaries of consciousness. Critics dispute whether those puzzles require a sharp fundamental element and whether the transition from microreality to an experiencing whole has been adequately explained. The account illustrates why a theory of character, a theory of conscious status, and a theory of subjects can come apart.

### Evidence and Open Disputes

Panpsychism does not gain empirical support merely because consciousness exists, nor is it refuted merely because simple entities cannot verbally report experience. Its contemporary arguments are largely metaphysical and comparative. The relevant question is whether positing a basic experiential nature explains more than rival accounts at an acceptable cost.

Critics press the absence of distinctive tests, the proliferation of subjects, and the relation of qualitative nature to physical causation. Defenders emphasize continuity, avoidance of brute emergence, and compatibility with physics. Claims that experiential properties must be causally idle do not reach views identifying them with the nature of causal powers; conversely, that identification still needs argument.

The principal unresolved issues concern constitution, individuation of subjects, the scope of physical explanation, and whether the argument from experience justifies a universal or only a local metaphysical conclusion.

### Further Reading


- Chalmers, David J. "Panpsychism and Panprotopsychism." *The Amherst Lecture in Philosophy* 8 (2013).

- Chalmers, David J. "The Combination Problem for Panpsychism." In *Panpsychism: Contemporary Perspectives*, edited by Godehard Brüntrup and Ludwig Jaskolla. Oxford University Press, 2017.
- Strawson, Galen. "Realistic Monism: Why Physicalism Entails Panpsychism." *Journal of Consciousness Studies* 13, nos. 10–11 (2006): 3–31.
- Tye, Michael. *Vagueness and the Evolution of Consciousness: Through the Looking Glass*. Oxford University Press, 2021.


**Related Writing on This Site**

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind and meaning.

**My View**

Two arguments need separating. Goff's Russellian route begins from physics's role/realizer gap and treats experience as positive *abductive* evidence—evidence in an inference to the best explanation—about categorical realization. Strawson's real physicalism begins from experience's physical reality and denies that experiential being could radically emerge from a wholly non-experiential base. Both reach fundamentality by different roads.

The strongest versions escape objections that defeat weaker ones. Panqualityism avoids micro-subject summing. Powerful qualities can make qualitative nature inseparable from causal power, blocking freely swappable or idle quiddities. The remaining move depends on an inheritance premise: organism-level manifest character must derive from qualitatively continuous fundamental properties rather than consist in relational organization at the level of the whole. Neither introspection, causal closure, nor structural physics establishes that premise.

Acquaintance may reveal a manifest phenomenal profile without revealing its absolute intrinsicness or fundamental ground. The book's local Russellian option can therefore disclose something of the organized physical episode while leaving its realizers open. If categorical character merely redescribes the concrete realization of a relational profile, the local view converges with the book. A necessarily co-instantiated realizer can remain metaphysically distinct even without a stable phenomenal contrast, but it gains no abductive advantage without independent support or explanatory payoff. A stable contrast with the complete independently specified content-undermode profile fixed would defeat the identity; its absence supports economy without proving equivalence.

The identity view pays its own price: a strong *a posteriori* necessity. The identity holds necessarily, on this view, but experience and inquiry rather than the concepts alone reveal it. One stopping rule applies to every package: evidence must exceed the bare existence of the target, unify a wider pattern or risk a discriminating failure, and beat rivals on fundamental commitments. On present evidence the book prefers identity because it explains phenomenal contrasts without adding phenomenal foundations or bridge laws, while displaying the result that would defeat it. The verdict remains categorical; its warrant remains comparative. See *Russellian Monism*, *Qualia*, and *Intentionalism*.

### The Phenomenal Concept Strategy

The phenomenal concept strategy explains the apparent gap between physical descriptions and conscious experience through distinctive ways of thinking about experience. Its physicalist proponents argue that phenomenal and physical concepts can refer to the same property without making that identity transparent to the thinker.

A scientist might describe a state through neural activity while a person undergoing it thinks "this kind of pain." The conceptual contrast, on the strategy, does not require two properties. Explaining why those concepts seem so different could weaken arguments that infer a metaphysical gap from an epistemic one.

The strategy is a family of proposals, not a universally accepted theory of concepts or a demonstrated physical explanation of consciousness.

### The Explanatory Task

An epistemic gap concerns what can be known, derived, or understood from a description. An ontological gap concerns what exists. Physicalists using the strategy accept some form of epistemic gap while denying that it establishes an ontological difference between experience and its physical realization.

The view is associated with *a posteriori* physicalism: some identities require experience and inquiry rather than following simply from meanings. Knowing a complete physical description need not, on these accounts, provide every way of thinking about what that description concerns.

Three tasks should be distinguished. A theory must explain what phenomenal concepts are, explain the epistemic effects attributed to them, and show that those effects can arise within a physical world. It must also have an independent reason to identify their referents with physical properties. Explaining the impression of a gap does not by itself prove the identity.

### Main Accounts

**Recognitional accounts.** Brian Loar treats phenomenal concepts as first-person recognitional or type-demonstrative concepts. A person can recognize a state as "that kind again" without being able to identify it through scientific theory. Their referents can be physical even though recognition does not work by deploying a physical description.

**Demonstrative and indexical accounts.** Other proposals emphasize expressions such as "this experience" or the subject's perspective. They compare phenomenal ignorance with knowing a fact descriptively while lacking its location relative to oneself. Critics ask whether a distinctive feel can be explained as a merely indexical limitation.

**Constitutional or quotational accounts.** On these proposals an experience, or a reactivated experiential representation, figures in the concept's vehicle. The contrast between thinking with such a vehicle and thinking through a theoretical description could explain the unusually direct character of phenomenal thought. A central objection asks whether the explanation presupposes the phenomenal nature it was supposed to reconcile with physicalism.

**Acquaintance-based accounts.** Some theories emphasize an epistemic relation to an occurring state. Acquaintance may ground later recognition without disclosing the state's complete metaphysical essence. These accounts differ over whether acquaintance is physically explicable and over how much it reveals.

The categories overlap, and not every acquaintance theory is physicalist. Likewise, recognizing special phenomenal concepts need not commit a philosopher to using them to explain away anti-physicalist arguments.

### Mary and Zombies

Applied to Mary's Room, the strategy permits a distinction between acquiring a new way of knowing and encountering a new nonphysical fact. Mary may gain a phenomenal concept of a state already covered by her physical knowledge. Some versions grant genuinely new propositional knowledge under that concept; others emphasize capacities rather than a new truth. They should not all be summarized as "Mary learns nothing."

Applied to zombies, the strategy explains why a thinker can apparently conceive physical duplication without experience. Physical and phenomenal concepts do not make their co-reference conceptually obvious. A critic replies that conceptual separability supports a deeper modal argument and cannot simply be labeled harmless. Whether conceivability licenses possibility remains an additional dispute.

### Chalmers's Dilemma

David Chalmers's "Phenomenal Concepts and the Explanatory Gap" challenges whether one set of psychological features can satisfy both explanatory demands. If those features cannot be explained physically, the gap reappears in the account of concepts. If they can be explained physically, a hypothetical physical duplicate without consciousness would share them; how, then, do they explain the conscious person's distinctive epistemic situation?

His Zombie Mary case focuses the issue. A zombie duplicate might produce the same reports and possess the same physically characterized recognitional machinery, yet lack Mary's phenomenal knowledge. An explanation of behavior and gap-talk seems weaker than an explanation of genuine knowledge.

Physicalist replies contest the possibility assumption, distinguish topic-neutral cognitive capacities from truth and reference, or argue that the strategy needs only to explain conceptual isolation once the psychophysical identity is independently supported. Chalmers questions whether that limited explanation sufficiently addresses the evidence motivating the anti-physicalist argument.

The dispute therefore concerns the required explanandum as well as the proposed mechanism. Explaining why people *say* the gap is deep and explaining why they possess phenomenal knowledge are not automatically the same task.

### Scope and Limits

A concept strategy should be distinguished from a theory of phenomenal character and from a theory of why a state is conscious. Representationalism, neural identity theories, and higher-order theories can supply different accounts of those questions.

Michael Tye's *Consciousness Revisited: Materialism without Phenomenal Concepts* (2009) explicitly rejects the need for a special phenomenal-concept explanation while remaining materialist. It is therefore inaccurate to classify all his work as advocacy for this strategy. Brian Loar's physicalist realism also shows that phenomenal-concept theories need not deny intrinsic phenomenal qualities.

Open issues include the distinctiveness of these concepts, the role of actual experience in their acquisition, their relation to memory, and the explanation of phenomenal knowledge. The strategy remains a substantive physicalist response with substantive objections, rather than a consensus resolution of the hard problem.

### Further Reading


- Loar, Brian. "Phenomenal States." *Philosophical Perspectives* 4 (1990): 81–108. Revised version in *The Nature of Consciousness*, MIT Press, 1997.
- Papineau, David. *Thinking about Consciousness*. Oxford University Press, 2002.
- Chalmers, David J. "Phenomenal Concepts and the Explanatory Gap." In *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, edited by Torin Alter and Sven Walter. Oxford University Press, 2007.
- Tye, Michael. *Consciousness Revisited: Materialism without Phenomenal Concepts*. MIT Press, 2009.


### Related Writing on This Site

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind and meaning.

### My View

The PCS supplies the dialectical setting for the book's epistemic-not-ontological framing, but the house account adopts no generic strategy wholesale. It pairs the independent identity claim with a narrower acquisition story: acquaintance with an occurring world-presenting state grounds recognitional capacities that description alone does not confer. See *Identity Theory*.

To Chalmers's dilemma the house account takes the second horn. A physical duplicate shares the architecture that anchors a demonstrative and later recognition; that architecture explains conceptual isolation and gap-talk, not by itself the truth of Mary's thought. Chapter 6′s independent identity claim fixes what the concept reaches and carries the metaphysical burden. Philip Goff adds a further question: what does acquaintance reveal? The answer grants limited translucency. Acquaintance reveals part of the state's manifest world- or body-presenting profile while leaving its content-fixing history and neural realization undisclosed. It reveals more than bare reference and less than complete essence.

### Phenomenology

Phenomenology studies structures of experience from the first-person perspective. The word also names a philosophical movement associated with Edmund Husserl and his successors. In contemporary philosophy of mind it has a looser use: "the phenomenology of pain" can simply mean what pain feels like, without endorsing Husserl's method or metaphysics.

A third use concerns *cognitive phenomenology*: whether conscious thinking has a distinctive experiential character beyond its accompanying sounds, images, and emotions. Keeping these uses separate helps avoid treating a description of experience as an automatic commitment to one philosophical school.

### Husserl's Method

Husserl investigates how objects appear in perception, memory, imagination, and judgment. The same tree can be seen from one side, remembered later, or imagined without being present. These acts differ in how they present their objects, not simply in which object is named.

*Intentionality*, the directedness of experience, supplies a central organizing idea. Husserl distinguishes the manner of an act from what is intended as intended. Interpretations of his technical distinction between *noesis* and *noema* differ, especially concerning whether the noema should be understood as meaning, an abstract entity, or the object under a particular mode of presentation.

The *epoché* suspends the ordinary assumption that the experienced world exists just as the natural attitude takes it to exist. It is not a declaration that the world is unreal. The point is to study the structures through which something is presented and accepted as real without relying unexamined on those assumptions.

Husserl's transcendental project goes beyond cataloging feelings. It asks about the conditions under which objects, knowledge, and a shared world become intelligible. His earlier descriptive work and later transcendental claims should not be collapsed into one doctrine of privately inspected mental objects.

### Time, Perspective, and the Body

Experience has temporal organization. Hearing a melody involves more than disconnected notes: the just-past note remains relevant while the next is anticipated. Phenomenological accounts examine this retention and anticipation without treating memory of a completed event as identical to the ongoing experience of a sequence.

Spatial perception is also perspectival. Seeing a cup from the front involves an implicit horizon of possible views and movements. Its hidden side can belong to what is anticipated without being visually presented in the same way as the visible surface.

Maurice Merleau-Ponty's *Phenomenology of Perception* emphasizes the lived body as a condition of such access. A staircase appears in relation to capacities for climbing; a tool can be incorporated into skilled movement. These descriptions challenge accounts that reduce perception either to passive sensory atoms or to detached intellectual judgments.

The phenomenological body is not merely a biological object observed from outside. This distinction between lived and objectively studied embodiment can guide empirical research without itself deciding whether bodily experience is reducible to physical processes.

### Variations Within the Tradition

Heidegger gives priority to practical involvement and being-in-the-world rather than treating detached observation as the model of all experience. Sartre and Merleau-Ponty develop different accounts of embodiment, freedom, and relations with others. Later phenomenology extends into social experience, affect, psychiatry, and ethics.

There is therefore no single phenomenological verdict on naturalism or on representational theories of mind. Some projects retain transcendental ambitions; others draw on descriptive distinctions while integrating psychology and neuroscience. Arguments about whether a theory can be naturalized must specify which project is under discussion.

### Phenomenal Character as an Explanandum

In analytic usage, phenomenal character names what undergoing a state is like. Seeing a bright light, feeling an ache, and understanding a sentence provide different candidate cases. This usage does not settle whether the character is intrinsic, relational, representational, or physical.

Representationalists appeal to properties presented in experience and to modes of presentation. Qualia realists allow character that exceeds representational content. Relational theories of perception emphasize the experienced object and its properties. Illusionists instead question whether the properties ordinarily ascribed to experience are accurately characterized.

These theories may share many experiential reports while disagreeing about their interpretation. Description is evidence for a theory; it does not become a metaphysical verdict simply by being first-person.

### Cognitive Phenomenology

David Pitt argues that conscious thought has proprietary, distinctive, and content-individuating phenomenology. Understanding a sentence, on this account, feels different from merely hearing an unfamiliar sequence even when much of the auditory input remains similar.

Opponents explain such differences through sensory imagery, attention, emotion, or a feeling that an utterance is comprehensible. They question whether introspection isolates a uniquely cognitive component. Defenders reply that matching those accompaniments can leave a difference in what is consciously understood.

This debate is not identical to asking whether inner speech exists. A person may think in words, but the words can support different interpretations, and some reported thought lacks verbal imagery. Jorba and Vicente investigate the relation between thought contents, conscious access, and inner speech rather than equating them.

### Objections, Replies, and Open Questions

A recurrent objection concerns disagreement. If careful reflection yields incompatible descriptions, what decides between them? Phenomenologists appeal to disciplined comparison, variation

of cases, and intersubjective scrutiny. Empirical approaches add experience sampling, behavioral measures, and clinical evidence. None provides an automatic, theory-free reading of experience.

Naturalistic critics question whether bracketing scientific assumptions can justify claims about the essential structure of consciousness. Defenders answer that causal explanation and clarification of how experience presents a world are different tasks. Disagreement persists over whether the tasks complement each other or impose conflicting explanatory priorities.

Transparency raises a further issue. Attending to seeing often directs attention toward the world seen, but reflection can also describe perspective, mode, and temporal organization. Whether those observations establish a representational, relational, or intrinsic-property account remains disputed.

Phenomenology therefore supplies methods, distinctions, and problems rather than an agreed solution to consciousness. Its contribution includes making the target of explanation more precise while leaving its metaphysical interpretation open.

**Further Reading**

- Husserl, Edmund. *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy, First Book*. Translated by F. Kersten. Martinus Nijhoff, 1983. German original, 1913.
- Merleau-Ponty, Maurice. *Phenomenology of Perception*. Translated by Colin Smith. Routledge & Kegan Paul, 1962. French original, 1945.
- Bayne, Tim, and Michelle Montague, editors. *Cognitive Phenomenology*. Oxford University Press, 2011.
- Smith, David Wood. "Phenomenology." *Stanford Encyclopedia of Philosophy*, revised 2013.

**Related Writing on This Site**

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind and meaning.

**My View**

Three legacies, each kept and reframed.

**From Brentano**: intentionality as the mark of the mental. The constitutive directedness of mental states distinguishes them from nonmental physical states without placing them outside the physical world. I take this as foundational, while replacing Brentano's immanent-object ontology with a causal-historical, externalist account of how intentional directedness gets fixed.

**From Husserl**: the discipline of attending carefully to the structures of experience as actually undergone. Any theory of mind must answer to phenomenology in the descriptive sense, however far it departs from the transcendental method — and I do depart from it. The structures phenomenology describes belong to embodied, world-engaged organisms, and they call for empirical-naturalistic explanation, not transcendental insulation.

**From Merleau-Ponty**: the embodied, world-involving character of perception and cognition. Perception is not the contemplation of inner representations by a disembodied observer; it is the active, bodily engagement of an organism with its environment. Embodiment as my view uses it (see *Embodiment*) inherits this directly.

On the contested questions above, I take sides. Phenomenology done well attends to the world as represented; phenomenology done badly tries to attend to an inner screen — and the transparency thesis, far from undermining the phenomenological project, tells you which of the two you are doing. On cognitive phenomenology my alliance is selective: there is something it is like to think that p, distinct from what it is like to see a sentence on a page or hear it spoken — Pitt is right about the datum. Where thought occurs in linguistic form, I read its phenomenal character as the phenomenology of *internalized public speech* (see *Inner Speech*): real, structured, world-directed, and inheriting its character from the public language in which it consists. Other modes of cognition — non-verbal problem-solving, expert performance, pre-linguistic infant thought — carry phenomenal character grounded in their own world-engaging activities. None of this requires non-physical inner properties or a sui generis "thought-stuff" alongside sensory phenomenology.

**Physicalism**

Physicalism holds that everything is physical or depends wholly on the physical in the appropriate way. In philosophy of mind, its central claim is that mental phenomena do not require an independent nonphysical basis. The thesis concerns what reality contains and how its levels relate; it does not say that ordinary psychological language is useless or that current physics already explains consciousness.

Defining both "physical" and "depends" creates substantial philosophical work. Physicalism is a family of metaphysical proposals rather than the slogan that minds are made from one familiar material.

**The Physical Base**

An older materialism pictured reality in terms of matter and motion. Contemporary physicalism need not limit its inventory to solid particles: fields and other entities described by physical theory can belong to the base.

*Hempel's dilemma* challenges appeals to physics. If "physical" means what current physics recognizes, the thesis may be hostage to an incomplete or mistaken theory. If it means whatever a future ideal physics recognizes, its content threatens to become indeterminate. Responses appeal to continuity with successful physics, the nature of paradigmatic physical objects, or exclusions such as the absence of fundamentally mental properties.

These formulations have different implications. Merely asserting one world does not distinguish physicalism from every form of property dualism. Nor does scientific investigability alone suffice: a theorist could regard nonphysical phenomenal properties as law-governed and investigable.

**Identity, Realization, and Grounding**

**Type-identity theories** identify kinds of mental state with kinds of physical state. A particular formulation might identify a kind of pain with a neural process. The often-used C-fiber example is a philosophical illustration, not a complete empirical theory of pain.

**Token physicalism** identifies each concrete mental event with a physical event. It is weaker than identifying mental kinds with one neural kind. By itself it may also leave open questions about the properties instantiated by those events.

**Realization theories** explain mental states through organized physical roles. A mental kind may be realizable in different physical systems, as the same computation can run on different devices. Multiple realizability motivates preserving a psychological level of explanation without accepting a separate mental substance.

**Grounding accounts** say that mental facts obtain entirely in virtue of physical facts. They try to articulate the dependence more strongly than simple covariation. What grounding adds to identity or realization, and whether it adequately explains consciousness, remain contested.

Reduction also has several meanings. Identifying a property with a physical property differs from translating every psychological statement into physical vocabulary or deriving every truth from a completed physics. Physicalists can disagree about explanatory reduction while agreeing on metaphysical dependence.

### Supervenience and Its Limits

A common formulation says that no mental difference is possible without a physical difference. Any appropriate physical duplicate must therefore be a mental duplicate. Strong versions concern metaphysical possibility, not just the actual laws governing familiar organisms.

Supervenience alone need not explain why the mental depends on the physical. A dualist can accept lawful psychophysical covariation, and an unrestricted duplication claim can face complications concerning additional nonphysical entities. More careful formulations restrict the comparison or add a "nothing over and above" condition.

Externalist theories also require an adequately wide physical base. If mental content partly depends on environment or history, duplication of a brain alone need not duplicate every mental fact. This does not itself challenge physicalism; it challenges an overly narrow specification of the physical facts.

### Arguments in Favor

Physicalists emphasize the explanatory success of physical science and systematic dependence of mentality on neural organization. Injury, stimulation, development, and disease change cognition and experience in patterned ways. Such evidence supports physical involvement, although dependence or correlation alone does not establish identity.

The causal argument adds physical causal completeness: physical effects have adequate physical causes. If mental states also cause bodily actions, identifying them with, or locating them within, those causes can avoid systematic overdetermination. Critics dispute completeness, the inference to identity, or the account of causation it presupposes.

For nonreductive physicalism this produces an internal challenge. If a realized mental property is distinct from its physically sufficient realizer, how does it exert causal influence without being excluded? Replies appeal to realization, shared causal powers, or differences between levels of causal explanation. Jaegwon Kim's work makes this exclusion problem central.

### Consciousness Objections

**The knowledge argument** asks whether Mary learns a nonphysical fact on first experiencing color despite complete physical knowledge. Replies distinguish new abilities, acquaintance, or modes of knowing from a new property.

**The zombie argument** asks whether a physical duplicate without experience is genuinely possible. If so, the relevant physicalism fails. Physicalists dispute the move from apparent conceivability to possibility or the conception of a complete duplicate.

**The explanatory gap** asks why the proposed physical realization has its particular experiential character. Some physicalists expect further empirical explanation; others appeal to features of phenomenal concepts or the limits of human understanding. Critics argue that these responses may explain ignorance without explaining consciousness.

Chalmers's taxonomy distinguishes type-A materialism, denying the relevant epistemic gap; type-B, accepting an epistemic but not an ontological gap; and type-C, anticipating a gap that can eventually close. These labels organize responses rather than establish their success. His dualist and Russellian categories mark other options, not a ranking automatically settled by the taxonomy.

### Boundaries and Open Issues

Physicalism does not logically require eliminativism about experience or belief. Eliminative theories reject some ordinary mental categories; reductive theories can preserve the phenomena while identifying them differently. Nor does physicalism alone establish substrate independence: some physicalists require particular biological properties, while others allow artificial realization.

Panpsychist and Russellian positions complicate the boundary. Some use "physicalism" while allowing experiential fundamental nature; others count such views as alternatives. The disagreement concerns what qualifies as physical, not simply how many worlds there are.

The thesis also differs from scientism, the claim that only scientific methods yield legitimate knowledge. Questions about abstract objects, values, explanation, and first-person knowledge remain relevant to formulations of a global physicalism. No single theory of consciousness follows merely from adopting the label.

### Further Reading


- Stoljar, Daniel. "Physicalism." *Stanford Encyclopedia of Philosophy*, revised 2026.
- Smart, J. J. C. "Sensations and Brain Processes." *The Philosophical Review* 68, no. 2 (1959): 141–156.
- Kim, Jaegwon. *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. MIT Press, 1998.
- Chalmers, David J. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, 1996.
- Chalmers, David J. "Consciousness and Its Place in Nature." In *The Blackwell Guide to Philosophy of Mind*, edited by Stephen P. Stich and Ted A. Warfield. Blackwell, 2003.

**Related Writing on This Site**

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind and meaning.

**My View**

Physicalism is the metaphysical position I keep arriving back at, even when I have tried hard not to. Other commitments have stayed open to revision — I have changed my mind on the identity claim and on how to engage higher-order theories — but on the metaphysics I have not been able to move. When an argument seems to force a dualist or panpsychist conclusion, my track record so far has been to find, on closer inspection, that the argument went wrong somewhere. That may say something about me or it may say something about the arguments, and I hold the position provisionally rather than as an article of faith.

My physicalism is **reductive but non-eliminative, relational, and embodied**, and it pairs a bare metaphysical floor with a positive theory. The floor: one world; everything in nature is, at bottom, physical; nothing mental is fundamental. The positive theory gives the floor its shape — strong representationalism (phenomenal character consists in representational content of the right embodied, world-directed kind), embodied realization, semantic externalism, and anti-reification. See *Intentionalism*, *Embodiment*, *Semantic Externalism*, and *Anti-Reification*.

**On the label, I lead with "physicalist" rather than "naturalist."** The word marks the one commitment the live rivals cannot share: phenomenal character is not fundamental. A property dualist will grant one world; a panpsychist will grant nature; both seat consciousness among the basic features of reality, and "naturalism" has gone roomy enough to shelter them. "Physicalist" closes that room. I hold the label provisionally — it reopens the lazy eliminativist misreading, so the surrounding articulation, not the bare word, has to carry the anti-eliminativist weight.

**In Chalmers's taxonomy, my view is Type B physicalism — and I will own the label plainly.** I accept the epistemic gap. Zombies remain conceivable; Mary gains genuine knowledge on release; the explanatory gap is real and is not dissolved by waving at the easy problems. That is where I part company with Type A. What I deny is the *ontological* gap. The psychophysical identities hold and are knowable only *a posteriori*. The broader phenomenal-concept literature maps several ways to explain the epistemic distance; the book adopts the narrower acquaintance-recognition account. Undergoing a world-presenting state grounds recognitional capacities that description alone does not confer. That explains the gap given the identity; Chapter 6′s independent case must establish the identity itself. See *The Phenomenal Concept Strategy* and *The Explanatory Gap*.

People usually put the question to me as: *how does matter become conscious?* I resist the phrasing, because it does its damage before any answer can begin. "Become" asks for two things and a transition between them — inert stuff on one side, an inner light on the other, and some moment where the first turns into the second. That is the picture I deny. The book instead separates four questions: what fixes content, which content-bearing states are conscious, which persisting organization bears them, and what world-presenting profile determines the character of states already conscious. Within that domain, phenomenal character and the relevant physical activity

amount to one fact reached from two directions. Practical stake may add autonomy and welfare; it supplies no ingredient of consciousness.

### Predictive Processing and the Interface

Predictive processing and the interface theory of perception both challenge simple pictures of perception as passive reception. They are nevertheless different frameworks. Predictive processing concerns how perceptual systems generate and revise hypotheses; interface theory concerns the relationship between adaptive perceptual representations and reality.

Neither the existence of prediction nor evolutionary selection by itself settles whether people directly perceive mind-independent objects. That philosophical question requires additional arguments about content, awareness, and what successful perception achieves.

### Prediction, Error, and Precision

Predictive-processing models describe a hierarchy in which higher levels predict lower-level activity and differences between prediction and input drive revision. Perception involves integrating sensory evidence with expectations rather than building a scene solely from an unstructured stream of signals.

A person trying to identify an indistinct figure through fog provides a useful illustration. Prior expectations can make one interpretation more likely, but a clearer view can overturn it. The balance depends on the reliability attributed to the evidence and the prior estimate. In technical accounts, *precision* concerns uncertainty; weighting by precision determines how strongly an error should influence revision.

Predictive coding is an algorithmic proposal for implementing this kind of updating. Bayesian descriptions can instead specify what an optimal estimate would be without claiming that the brain literally executes one particular algorithm. Distinguishing the computational aim, algorithm, and neural implementation prevents empirical success at one level from proving every stronger claim.

The prediction need not be a conscious expectation or an explicit proposition. Much of the proposed processing is subpersonal. Describing a mechanism as drawing an inference does not mean that a little scientist inside the head reasons from premises.

### Perception and Action

Action can change the incoming evidence. Moving closer to a shape, turning an object, or directing the eyes helps reduce uncertainty. Active-inference approaches integrate this sampling with the prediction framework: the organism changes its situation as well as updating its estimates.

Andy Clark emphasizes embodied and environmentally situated prediction. Bodies, tools, and cultural arrangements can simplify the problems an agent must solve. Jakob Hohwy emphasizes the inferential and evidentiary boundaries of the predictive system. Their accounts differ in philosophical interpretation and emphasis; neither should be reduced to a slogan about disembodied brains or automatic openness to the world.

The free-energy principle is a related, broader proposal about adaptive systems. It is not simply another name for every predictive-processing model. Claims of unification need to be distinguished from specific experimentally testable claims about perceptual circuitry.

### "Controlled Hallucination"

The phrase "controlled hallucination," prominent in discussions by Anil Seth and also used in predictive-processing debates, stresses that both ordinary perception and hallucination involve internally generated expectations. In ordinary perception sensory evidence constrains the result more successfully.

The phrase can mislead if it suggests that successful perception is clinically hallucinatory or that an internal model is necessarily the object of awareness. Generating a presentation does not by itself establish what the subject perceives. A theory of neural processing and a theory of the objects of awareness address different questions.

Nico Orlandi and Geoff Lee further challenge excessively top-down descriptions of the framework. Hierarchical models can integrate strong sensory evidence, and current input may dominate expectations. The explanatory claim should specify the circumstances and mechanism rather than assume that input contributes little to every percept.

### Hoffman's Interface Theory

Donald Hoffman, Manish Singh, and Chetan Prakash defend a more radical claim about perceptual evolution. Their models compare perceptual strategies tuned to objective structure with strategies tuned to fitness payoffs. Under modeled conditions, payoff-oriented strategies can outcompete veridical ones.

The desktop analogy expresses the proposed consequence. A file icon guides useful action without resembling the circuitry that implements the file. Similarly, a perceptual world of objects in space and time could guide adaptive behavior without reproducing reality's fundamental structure.

The theory does not say the icon is harmless or that appearances may be ignored. A speeding car should still be taken seriously within the interface. The claim concerns whether its perceived spatial and object structure reveals objective reality. It is stronger than the familiar observation that an organism notices only selected environmental features.

### Objections and Replies

**Fitness and tracking.** Critics argue that successful action often requires reliable sensitivity to environmental structure. Interface theorists reply that payoff-relevant sensitivity need not preserve that structure in the way a realist expects. The debate must specify what counts as truth-tracking: exhaustive resemblance, selective structural preservation, and reliable categorization are different standards.

**Model scope.** Evolutionary simulations depend on assumptions about worlds, fitness functions, costs, and available strategies. A result for those models does not directly establish that human perception is globally nonveridical. Advocates argue that the models expose a general mistake

in expecting selection to favor truth; critics dispute the transfer from simplified competitors to biological perception.

**Scientific evidence.** A radical interface theory uses perceived experimental results to support a theory that questions ordinary perceived structure. Critics see a threat to its evidential basis. Defenders distinguish practical and predictive success within an interface from fundamental resemblance and argue that science can infer beyond appearances. Whether that distinction supplies adequate warrant remains contested.

**Prediction's empirical reach.** Evidence for expectation effects does not establish one universal predictive architecture. Competing mechanisms can sometimes produce similar behavior, and flexible models risk accommodating too many outcomes. More specific predictions about timing, circuitry, and response to intervention offer stronger tests than a broad claim that the brain anticipates.

### Philosophical Options

A representational direct realist can accept predictive mechanisms while holding that their content concerns external objects. An indirect realist may regard the subject as aware of an internally constructed intermediary. A relational theorist can accept extensive processing while denying that the resulting experience is fundamentally a representation.

The vehicle-content distinction helps keep these alternatives intelligible, but does not by itself choose among them. The open questions concern the architecture of perception, the warrant for evolutionary extrapolation, and whether an inferential account explains or changes the ordinary claim to perceive the world.

### Further Reading


- Hohwy, Jakob. *The Predictive Mind*. Oxford University Press, 2013.
- Clark, Andy. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press, 2016.
- Orlandi, Nico, and Geoff Lee. "How Radical Is Predictive Processing?" In *Andy Clark and His Critics*, edited by Matteo Colombo, Elizabeth Irvine, and Mog Stapleton, 206–221. Oxford University Press, 2019. doi:10.1093/oso/9780190662813.003.0016.
- Hoffman, Donald D., Manish Singh, and Chetan Prakash. "The Interface Theory of Perception." *Psychonomic Bulletin & Review* 22 (2015): 1480–1506.


### Related Writing on This Site

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind and meaning.

### My View

A representational, prediction-driven brain is exactly the kind of brain a direct realist should expect. Representation is *how* the organism reaches the world, not a screen between them, and the vehicle/content distinction settles the question for me: the model is the seeing, not the seen. I welcome the science without reservation. The veil gloss is a philosophical add-on the science never earned — a metaphysics smuggled in under a slogan. Introspect the experience and you find

the world, not the model, which is what the transparency of experience predicts and the interface picture has to explain away.

### Qualia

Qualia, singular *quale*, are the qualitative aspects of experience: what seeing a color, tasting coffee, or feeling pain is like. The word has several uses. In a broad use it names phenomenal character without deciding what that character consists in. In narrower uses it denotes intrinsic or nonrepresentational properties of experiences. Disputes about whether qualia exist often turn on which use is intended.

A person can affirm that pain genuinely hurts while denying one philosophical account of that hurt. Conversely, a theory can allow intrinsic phenomenal properties without treating them as objects watched by an inner observer. Definition should not settle the metaphysics before the arguments begin.

### Broad and Restrictive Conceptions

The broad conception identifies qualia with experiential character. It leaves open whether that character is physical, representational, relational, or fundamental. An account of color experience and an account of bodily pain need not assign the qualities to the same kind of thing.

A more restrictive conception treats qualia as intrinsic qualities of the experiential state rather than properties represented in its content. Still stronger packages add ineffability, privacy, or direct and infallible introspective access. These claims are independent enough to require separate defense.

Daniel Dennett's critique of qualia targets a restrictive cluster of commitments. It should not be assumed that every realist accepts that cluster. Ned Block's "Mental Paint," for example, defends character beyond the intentional, cognitive, and functional while allowing the prospect of empirical reduction. Realism about qualia therefore does not entail dualism.

### The Knowledge Argument and Explanatory Gap

Frank Jackson's Mary knows every physical fact about color vision but has never experienced color. If she learns what seeing red is like on release, does she discover a nonphysical fact? The argument supports an anti-physicalist conclusion only if her gain cannot be understood as new abilities, acquaintance, or a new way of knowing a fact already covered by physical knowledge.

Joseph Levine's explanatory gap asks why a particular physical process should have its particular feel. Even a proposed identity can leave that question pressing. The existence of an explanatory gap does not settle whether it reflects our concepts, a limit of explanation, or a metaphysical difference.

Jackson later rejected his original anti-physicalist conclusion. Levine's gap should not simply be assigned the same conclusion as the knowledge argument. The arguments motivate related concerns but have distinct premises and possible responses.

### Inverted and Absent Qualia

An inverted-spectrum scenario asks whether two people could have systematically different color experiences despite comparable discrimination, naming, and action. Someone who calls grass green might experience what another person experiences when seeing red, while both have learned the same public vocabulary.

The case challenges a simple inference from behavior to phenomenal sameness. To challenge a particular functional or representational theory, however, it must hold fixed the organization or content that theory specifies. Real color experience also has asymmetries of similarity and discrimination that make a perfect, behaviorally undetectable inversion harder to specify than swapping two labels.

An absent-qualia scenario instead asks whether the relevant functional organization could occur without any experience. Block's objections to functionalism explore cases intended to separate causal organization from phenomenality. Defenders of functionalism question the intuitions or argue that genuine duplication must include more than superficial input-output equivalence.

Externalist representationalists also ask whether environment and history are fixed. Their reply cannot simply count as a refutation: critics challenge whether wider content explains felt character and whether the enlarged specification has independent justification.

### Representational and Relational Accounts

Strong representationalists identify phenomenal character with content of an appropriate kind. A red-looking tomato experience presents redness as a property of the tomato, not of the neural process. A matching hallucination can present the same property without an object satisfying the presentation.

Transparency supplies a motivation: attending to the visual quality tends to direct attention toward the presented scene. Critics, including Block, argue that this does not exclude an intrinsic quality contributing to presentation. Representationalists dispute the need for such an additional quality and develop content-based explanations of inversion, pain, and perceptual organization.

Relational theories of successful perception instead emphasize the actual object and its properties as constituents of the experience. Their treatment of hallucination may be disjunctive: an indistinguishable bad case need not share the good case's fundamental nature. This is another alternative to intrinsic qualia, not merely a version of content-based reduction.

### Access and Phenomenality

Block distinguishes *access consciousness*, availability for reasoning and the control of action, from *phenomenal consciousness*, there being something it is like to undergo a state. The distinction separates concepts; whether the corresponding phenomena can come apart is an additional question.

Overflow arguments suggest that experience contains more than a person can access for report. Critics attribute the excess to unconscious representation, memory, or an overconfident judgment about detail. Blindsight, in which subjects can discriminate stimuli they deny seeing, raises complementary questions about access, degraded experience, and what a report establishes.

None of these findings straightforwardly proves intrinsic mental paint. A representational theory of phenomenal character can recognize the conceptual distinction while needing its own account of conscious status. Likewise, accepting access mechanisms does not by itself answer what makes a state feel as it does.

### Illusionism and Introspection

Keith Frankish distinguishes **weak illusionism**, which retains genuinely qualitative properties while rejecting alleged features such as ineffability or infallible access, from **strong illusionism**, which denies genuinely qualitative what-it-is-like properties, whether physical or nonphysical. Merely rejecting the restrictive qualia package need not amount to strong illusionism: a representationalist can reject that package while affirming phenomenal character.

Strong illusionists argue that introspection misrepresents sensory states as having phenomenal properties. They seek to explain sensory processing, functional pain, discrimination, and introspective judgments without such properties. Their claim is stronger than the familiar observation that introspection sometimes makes mistakes.

Critics ask whether an illusion itself presupposes conscious seeming and whether the explanatory account preserves what needed explaining. Illusionists distinguish the cognitive disposition to judge that a special feel is present from possession of that feel. The disagreement should not be compressed into the claim that pain is unimportant or that nothing happens when someone suffers.

Introspection provides evidence but remains disputed in scope and reliability. Recognizing experience, identifying its causes, and knowing its metaphysical nature are different achievements. A complete account of qualia must specify which conception it accepts and explain why that conception fits the evidence better than alternatives.

### Further Reading


- Tye, Michael. "Qualia." *Stanford Encyclopedia of Philosophy*, revised 2025.
- Block, Ned. "Mental Paint." In *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, edited by Martin Hahn and Bjørn Ramberg. MIT Press, 2003.
- Block, Ned. "On a Confusion about a Function of Consciousness." *Behavioral and Brain Sciences* 18, no. 2 (1995): 227–247.
- Levine, Joseph. "Materialism and Qualia: The Explanatory Gap." *Pacific Philosophical Quarterly* 64 (1983): 354–361.
- Frankish, Keith. "Illusionism as a Theory of Consciousness." *Journal of Consciousness Studies* 23, nos. 11–12 (2016): 11–39.


### Related Writing on This Site

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind and meaning.

### My View

Qualia as traditionally conceived — intrinsic, non-relational, directly accessible inner properties of experience — do not exist. The concept involves a systematic reification. It treats the way experience represents the world as an inner substance the experience *has*, rather than a relation

it *stands in*. Phenomenal character is real; the inner-object interpretation of it is not. Phenomenal character consists in representational content of the right world-involving kind — relational through and through.

Behind the realist picture stands an implicit model: experience presenting inner objects (qualia) to an inner observer (the subject), with "mental paint" spread across a private inner surface. The transparency argument targets exactly that model. Attend to your experience of the tomato and you find the tomato — its redness, its sheen — never an inner surface, an inner observer, or inner paint. On my diagnosis, the inner-theater picture is a philosophical projection onto experience, not a discovery within it. Once the theater goes, qualia as traditionally conceived go with it, and what remains is what was really there all along: world-presenting phenomenal character, recast as representational content, fully inside the physical world. Transparency and strong representationalism make the positive case; Sellars supplies the structural diagnosis of why the realist picture never cohered.

### Qualia Realism

Qualia realism affirms the reality of qualitative features of experience. Its content depends on what "qualia" means. In a broad sense, many rival theories agree that experiences have phenomenal character. In a stronger sense, realism holds that character is not exhausted by representational or functional facts. It is the stronger claim that generates much of the contemporary dispute.

A realist can regard phenomenal properties as intrinsic properties of physical experiential states. Alternatively, a realist can regard them as nonphysical or fundamental. Neither version requires a mental picture observed by a second, inner subject, and neither follows merely from saying that pain hurts.

### What a Substantive Realist Claims

Seeing red differs phenomenally from seeing blue, and pain differs from visual experience. A substantive realist argues that these differences require qualitative properties not captured wholly by what the state represents or by its causal role.

The view need not deny representation. Ned Block argues that sensations typically, perhaps always, have representational content, and that their character may itself bear that content. His disagreement concerns whether content exhausts the character. "Mental paint" names the disputed contribution; it need not mean an inner surface literally painted red.

Intrinsicness is also a separate claim from infallible access. A property can be intrinsic to an experiential state while a subject's judgments about it are incomplete or mistaken. Realists disagree over what introspection reveals and whether apparent privacy entails anything beyond a special first-person epistemic route.

### Physicalist and Anti-Physicalist Versions

**Physicalist realism** allows phenomenal properties to be identical with, or wholly realized by, physical properties. It can reject philosophical reduction to functional or representational descriptions while leaving empirical identification open. Block's "Mental Paint" expressly distinguishes those kinds of reduction.

Sydney Shoemaker's work explores qualitative character in connection with causal and functional roles. It should not be grouped indiscriminately with theories that make qualia wholly independent of function. Realism is compatible with substantial disagreements about the causal individuation of qualities and the possibility of absent qualia.

**Dualist realism** treats at least some phenomenal properties as additional to the physical base. David Chalmers's arguments use explanatory and modal considerations to motivate such a position. Affirming phenomenal reality alone is weaker than accepting this conclusion: the inference needs premises about what physical explanation and duplication can supply.

**Panpsychist and Russellian versions** locate phenomenal or related qualitative nature within reality's fundamentals. They aim to avoid a wholly nonexperiential base but face constitution and combination questions. These positions should not be equated with every form of realism about ordinary experience.

### Arguments and Physicalist Replies

The **knowledge argument** asks whether Mary acquires knowledge of a nonphysical fact on first seeing color. It motivates realism by emphasizing something a complete third-person account appears to leave unavailable. Ability, acquaintance, and phenomenal-concept replies explain her epistemic gain without accepting an extra property. Realists ask whether these replies account for what is learned or simply assume a physical referent.

The **explanatory gap** asks why a particular physical organization has its particular character. Some realists regard this as evidence against reduction; others are physicalists who think the identity can be true without being transparent to explanation. The gap therefore does not uniquely support dualism.

**Inversion and absent-qualia cases** challenge the sufficiency of function or representation. Defenders must specify what remains fixed, since identical behavior need not imply identical functional organization or wide content. Opponents of the thought experiments ask whether their apparent possibility survives a complete specification. The disagreement is modal as well as empirical.

**Zombie arguments** target physicalism more directly by proposing a complete physical duplicate without experience. Realism about qualitative properties does not require accepting this possibility. A physicalist realist can hold that the properties are necessary consequences of, or identical with, physical properties despite conceptual separability.

### Transparency and the Mental-Paint Dispute

Representationalists argue that attention to visual experience reveals colors and shapes as features of a presented scene. A tomato appears red; the neural process need not itself be red. They interpret this transparency as evidence that phenomenal character belongs to the representational presentation rather than an additional intrinsic quality.

Realists reply that an experience can direct attention outward while having an intrinsic property partly constituting how the world appears. Outward attention does not by itself reveal the

metaphysical basis of presentation. An internalist theory can also permit hallucination to share a character with seeing even when environmental relations differ.

The representationalist response asks for a demonstrable contrast that their independently specified content-and-mode account cannot explain. Realists challenge whether such specifications are genuinely independent or expand to absorb each counterexample. Both theories need to distinguish explanatory content from an inventory of everything reported about a feel.

Neural differences are not automatically discriminating evidence: representational properties also require physical realization. A useful empirical contrast would need to connect a specific realization difference with character while controlling the intentional features the rival theory claims are sufficient.

### Illusionism and the Burden of Explanation

**Strong illusionism**, as defended by Keith Frankish, denies genuinely qualitative what-it-is-like properties, whether physical or nonphysical, and seeks to explain why people judge that they have them. It retains sensory processing and functional pain while denying their alleged phenomenal character. Frankish distinguishes this from **weak illusionism**, which preserves qualitative properties but rejects some supposed features, such as ineffability or infallible access. Rejecting those features alone can therefore remain a form of phenomenal realism.

Realists argue that strong illusionism substitutes an explanation of judgment for an explanation of experience. Strong illusionists reply that the target itself has been misconceived and that a successful account of introspective judgments can dissolve the demand.

The dispute cannot be resolved by changing labels. Saying that character is "real" requires a theory of what makes it real; denying "qualia" requires specifying what is denied. A realist need not accept ineffability, an epistemic Given, or an inner theater, just as a representationalist need not deny experience.

Open issues concern intrinsicness, realization, conscious access, and what introspective or neuroscientific evidence could distinguish the competing explanations. These issues leave room for physicalist as well as anti-physicalist realism without making their arguments equally successful by definition.

### Further Reading


- Block, Ned. "Mental Paint." In *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, edited by Martin Hahn and Bjørn Ramberg. MIT Press, 2003.
- Shoemaker, Sydney. "Phenomenal Character." *Noûs* 28, no. 1 (1994): 21–38.
- Chalmers, David J. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, 1996.
- Levine, Joseph. "Materialism and Qualia: The Explanatory Gap." *Pacific Philosophical Quarterly* 64 (1983): 354–361.
- Tye, Michael. "Qualia." *Stanford Encyclopedia of Philosophy*, revised 2025.
- Frankish, Keith. "Illusionism as a Theory of Consciousness." *Journal of Consciousness Studies* 23, nos. 11–12 (2016): 11–39.

**Related Writing on This Site**

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind and meaning.

**My View**

I reject qualia realism, but carefully, and on the representationalist rather than the eliminativist side of the divide. The realist is right that experience has phenomenal character — there really is something it is like to see red. Some versions then reify that character as an inner item. Block's strongest physicalist version need not: it can assign worldly accuracy to content and phenomenal constitution to an intrinsic physical property of the neural vehicle. The case against it must therefore remain comparative. Across the cases examined, independently measured changes in content, mode, determinacy, field structure, and attention track the phenomenal contrasts, while no independently identified vehicle property predicts a contrast the complete profile misses. A stable difference with subject and profile fixed would reverse that verdict. See Vehicle and Content and Transparency.

So I do not say, with the eliminativist, that there are no qualia in *any* sense. I say that qualia *as the realist conceives them* — intrinsic, non-representational inner items — do not exist. What exists is phenomenal character: within a conscious state, the complete world-presenting intentional profile — content under a perceptual or bodily-affective mode, with determinacy, field structure, and attentional organization. The qualities presented in experience belong in the first instance to surfaces, objects, and bodily conditions, not to an inner surface. See Qualia and Intentionalism.

The arguments that motivate realism do real work, and I take them seriously rather than waving them off. Mary gains a new representational capacity, not a new non-physical fact. The explanatory gap sits in our concepts and explanatory practices, not in nature. Inversion, absent qualia, and physicalist mental paint remain direct tests of the identity claim; the book answers them by specifying the complete profile independently and posting the evidence that would pull profile and character apart. See The Explanatory Gap and Mary's Room.

This matters because qualia realism, in its dualist form, is where the hard problem comes from. If phenomenal properties are real and intrinsic and irreducible, the problem of explaining how they arise from the physical becomes genuinely insoluble — by construction. Deny the realist's intrinsic inner property without denying the phenomenal character, and the problem deflates rather than remaining to be solved. The identity claim, not particle physics, is what closes the gap. See The Hard Problem and Physicalism.

**Reduction and Elimination**

Reduction and elimination offer different verdicts about a phenomenon. A reduction identifies it with, or explains it through, something more basic while retaining its existence. Elimination denies that the relevant category successfully describes anything and replaces the theory that introduced it. Water's identification with $H_2O$ preserved water; the rejection of phlogiston removed a supposed substance from chemistry.

In philosophy of mind, the distinction concerns such things as beliefs, desires, pains, and phenomenal qualities. A physicalist can defend some of these categories while rejecting others.

Physicalism alone therefore does not decide between reduction and elimination, and the alternatives need not exhaust the available forms of scientific revision.

### Forms of Reduction

**Ontological reduction** concerns what exists: a mental state or property may be identical with a physical state or property. **Explanatory reduction** concerns explanations or theories: neuroscience may explain a psychological regularity through lower-level mechanisms. **Conceptual reduction** analyzes a concept in other terms, potentially allowing conclusions to follow a priori, without further empirical discovery. As Tim Crane emphasizes in *Elements of Mind*, ontological and explanatory reduction can come apart.

Water and lightning illustrate empirically discovered identities, not proofs that every mental phenomenon admits an analogous treatment. Even these scientific examples require qualification. The relation between temperature and molecular motion depends on the system and statistical assumptions; a slogan about mean kinetic energy does not capture every use of temperature.

An identity claim also differs from a realization claim. Saying that a particular neural process realizes a functional capacity need not identify the capacity's general type with that neural type. Different materials might realize the same capacity. The distinction underlies multiple-realizability objections to simple mind–brain type identities and supports versions of nonreductive physicalism.

### What Eliminativism Targets

Paul Churchland's "Eliminative Materialism and the Propositional Attitudes" treats ordinary belief-and-desire psychology as a theory whose central categories might fail to survive mature neuroscience. Its motivation comes partly from the history of unsuccessful theories and partly from alleged limits of folk psychology in explaining learning, sleep, and mental illness.

This is not a prediction that people will stop speaking, acting, or reasoning. The proposed elimination concerns the classification of those activities in terms of beliefs and desires. Critics respond that ordinary psychology has substantial explanatory success and may impose fewer theoretical commitments than the eliminativist supposes. A future neuroscience might explain why belief-talk works rather than show that no beliefs exist.

Debates about qualia require a separate distinction. Some philosophers reject qualia understood as intrinsic, ineffable, privately accessible properties while retaining experiences and their phenomenal character. Others argue that phenomenal consciousness, as ordinarily conceived, involves a systematic illusion. Rejecting one technical definition of *qualia* does not by itself amount to denying every sense of conscious experience.

### Revision Between the Extremes

Scientific change often revises categories rather than either retaining them intact or eliminating them wholesale. A psychological term might cover several neural kinds; some aspects of its original theory might survive while others disappear. William Ramsey's survey of eliminative materialism describes reduction and elimination as endpoints with revisionary possibilities between them.

The examples of water and phlogiston therefore establish a useful contrast, not a mechanical test. Deciding which pattern applies requires asking which features define the category, which examples genuinely instantiate it, and how much alteration its continued use can tolerate.

**Conscious Experience and the Explanatory Gap**

Michael Tye's representationalism identifies phenomenal character with representational content satisfying further conditions. It aims to explain experience, rather than eliminate it. Critics nevertheless ask whether the proposed content preserves what seeing red or feeling pain is like, or substitutes a different explanatory target. A theory's declaration that it preserves experience does not settle that objection.

Joseph Levine's explanatory gap introduces another complication. A physical identity might be true without our being able to explain why the physical state has that particular phenomenal character. Physicalists may treat the gap as conceptual or epistemic; opponents may take it as evidence against the proposed reduction. Neither an unexplained identity nor the rejection of a priori deduction automatically amounts to elimination.

Open questions concern the criteria for successful reduction, the role of multiple realization, and the extent to which familiar mental categories can change while retaining their referents.

**Further Reading**

- Churchland, Paul M. 1981. "Eliminative Materialism and the Propositional Attitudes." *Journal of Philosophy* 78: 67–90.
- Crane, Tim. 2001. *Elements of Mind: An Introduction to the Philosophy of Mind*. Oxford University Press.
- Levine, Joseph. 1983. "Materialism and Qualia: The Explanatory Gap." *Pacific Philosophical Quarterly* 64: 354–361.
- Tye, Michael. 1995. *Ten Problems of Consciousness*. MIT Press.

**Related Writing on This Site**

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind, matter, and meaning.

**My View**

The charge I most want to answer squarely is that my view is eliminativism wearing a nicer suit — that once I say phenomenal character *consists in* representational content, I have quietly deleted the feeling and kept the word. It is a fair thing to press, and I have felt its force, so let me meet it at full strength before I answer it.

Here is the objection at its best. If the redness of red just *is* a representational state, then all the explanatory work is being done by the representing, and the felt quality is along for the ride — a name we keep out of sentiment for something the theory has already dissolved. That would indeed be elimination in disguise, and I would want no part of it.

But it mistakes what a reduction does. I am not saying there is nothing it is like to see red. I am saying there *is* something it is like, that this is a genuine datum no theory may deny, and that the

identity claim is my account of what that something *is* — not my excuse for changing the subject. A simple test: an eliminativist about an X denies that there are any Xs. I affirm that there are experiences, that they have felt character, that the character is exactly as rich as it seems. What I deny is a *picture* of that character — the picture on which it is a private inner object, a patch of mental paint, a further fact floating above the world-presenting activity. Denying the picture is not denying the phenomenon. When we learned that lightning consists in electrical discharge, we did not learn that the sky had been empty all along.

The sense of "reduction" I mean is the ontological one, not the conceptual: no armchair deduction runs from the physics to the feeling, and I concede that gap as fully epistemic — while insisting the identity holds all the same.

So I hold the line on both sides at once. Against the dualist and the qualia realist: phenomenal character is not a fundamental, free-standing ingredient of reality; it consists in something more basic. Against the eliminativist: it is entirely real, preserved, not explained away — the reduction is *of* it, and would be false if there were nothing there to reduce. Reductive but not eliminative is not a slogan I use to have it both ways. It is the only position that takes the feeling seriously *and* refuses to make it a miracle. The eliminativist saves the metaphysics by sacrificing the datum. The dualist saves the datum by sacrificing the metaphysics. I decline the trade. The feeling is real; here is what it consists in; both halves are meant.

**Reference**

Reference concerns how an expression, representation, or speaker picks out something. A name may refer to a person, a demonstrative to an object in the present situation, and a general term to a kind. The relation looks simple when someone points to an apple and says "that." Its complexity appears when speakers disagree about the object, inherit a name through testimony, or use a meaningful expression with no existing bearer.

Reference is one aspect of semantics, not an uncontested theory of all meaning. Logical connectives, commands, fictional discourse, and descriptions pose different problems from ordinary proper names. Philosophers also distinguish the reference associated with an expression from what a speaker intends to refer to on a particular occasion.

**Sense, Reference, and Empty Expressions**

Gottlob Frege's distinction between *sense* and *reference* addresses a familiar puzzle. "Hesperus is Phosphorus" can convey an astronomical discovery, whereas "Hesperus is Hesperus" ordinarily does not, although both names designate Venus. Frege explains the contrast through different modes of presentation of the same referent. Reference alone does not capture the difference in cognitive significance.

Bertrand Russell's "On Denoting" analyzes definite descriptions rather than treating every meaningful phrase as a name. "The present king of France is bald" receives a quantified analysis requiring a unique present king of France. If there is none, the sentence fails that requirement; a nonexistent king need not become an object in the theory. Russell's treatment of descriptions should not be conflated with a single solution to every empty proper name, such as "Pegasus." Fictional names remain a separate and disputed topic.

Direct-reference theorists can agree that two sentences differ in cognitive significance while denying that a Fregean sense forms part of a name's semantic contribution. The Hesperus example motivates a problem; it does not settle which theory must solve it.

### Descriptive and Causal-Historical Accounts

Descriptivist theories connect a name's reference with descriptions associated with its bearer. A name might refer to whoever uniquely satisfies an identifying description or an appropriate cluster of descriptions. This promises to explain how a speaker locates the bearer and why an apparently empty name can remain intelligible.

Saul Kripke challenges descriptive accounts through cases of ignorance and error. A speaker can refer to someone despite knowing very little about that person or associating the name with a false description. His causal-historical picture traces reference through an initial fixing and subsequent transmission in which speakers intend to preserve the earlier use.

Kripke also argues that ordinary names function as **rigid designators**: they designate the same individual across possible circumstances in which that individual exists. Rigidity, reference fixing, and direct reference are distinct claims. A description can fix the bearer of a name without supplying its meaning; a causal history is a proposal about how reference persists, not simply a synonym for rigidity.

### Context and Speaker Reference

Indexicals such as "I," "here," and "now" depend on features of their context. Demonstratives such as "that apple" may additionally depend on attention, gesture, and intention. A dictionary definition alone cannot identify the object of each use.

Speaker reference can also diverge from an expression's conventional reference. Someone who says "the man drinking a martini" may successfully draw attention to a man whose glass actually contains water. Whether the description itself refers, whether the speaker refers, and what the utterance says require separate analyses. This distinction motivates disagreements about semantics and pragmatics rather than a general conclusion that descriptions have collapsed as referring devices.

### Objections and Unresolved Questions

Descriptive theories face cases where speakers refer despite false or incomplete identifying beliefs. Causal-historical theories face the converse difficulty: many causal chains connect a word with objects it does not designate. They must explain which transmission counts and how reference can change through mistakes or evolving practice. Hybrid theories combine historical continuity with descriptive or intentional constraints.

Externalist accounts emphasize environmental and social relations. Their critics ask how those relations explain speakers' understanding, cognitive differences, and apparent first-person knowledge of what they mean. The claim that every meaningful practice ultimately requires reference to an independently existing world goes beyond the bare definition of reference and remains a substantive philosophical thesis.

Applications to artificial intelligence raise related questions about inherited linguistic reference, internal representation, and successful acts of referring. A system's use of public names, its representation of a specific object, and its understanding of an utterance need not receive the same answer.

**Further Reading**

- Frege, Gottlob. 1892. "On Sense and Reference." Translated in *Translations from the Philosophical Writings of Gottlob Frege*, edited by Peter Geach and Max Black. Blackwell, 1952.
- Russell, Bertrand. 1905. "On Denoting." *Mind* 14: 479–493.
- Kripke, Saul A. 1980. *Naming and Necessity*. Harvard University Press.
- Michaelson, Eliot. 2024 revision. "Reference." *Stanford Encyclopedia of Philosophy*.

**Related Writing on This Site**

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind, matter, and meaning.

**My View**

The collapse of the referential theory teaches a precise lesson, and the lesson is not that reference stopped mattering. Meaning *requires* reference. The requirement operates at the level of a linguistic practice; it does not assign successful denotation to every meaningful expression. Reference supplies a necessary ground of linguistic meaning, not the universal semantic function of every word. What collapses is the picture of reference as a label, a relation legible off the symbol. Reference survives as a world-embedded achievement: a causal-historical, community-sustained reaching from word to world.

Meaning-as-use and reference are not rivals. Use becomes semantic within practices sustained by **world-directed intentional systems** whose perceptions, corrections, inferences, and actions answer to something beyond the signs. Reference gives those practices their worldward anchor and keeps use from collapsing into mere regularity among symbols. A text-trained mechanism can carry derived content and inherit public reference through the practice that produced its corpus. Training alone does not fix a distal singular address. A system's own concrete *de re* reference—or its own asserting, promising, or asking—requires a history and organization that earn the relevant world-directed intentional role, plus consumer integration sufficient to attribute the act to that system. Practical stake bears separately on autonomy and welfare.

**Russellian Monism**

Russellian monism names a family of theories built from an insight Bertrand Russell developed most clearly in *The Analysis of Matter* (1927). Physics may tell us the structure of matter—its relations and causal powers—without telling us its complete concrete nature. Contemporary Russellian monists join that limit to a proposal about consciousness. The same underlying properties, they argue, may ground both physical causation and phenomenal character.

The modern family extends Russell's insight; it should not be presented as one doctrine Russell himself cleanly endorsed. Chalmers, Strawson, Goff, and others have developed several metaphysically different descendants under the Russellian name.

### The Core Argument

The argument has three stages.

1. **Structural physics.** Fundamental physics characterizes properties through equations, relations, dispositions, and causal roles. Mass gets characterized through resistance to acceleration and gravitational behavior; charge through attraction and repulsion.
2. **Acquaintance.** Experience gives a subject first-person access unavailable from structural description. I do not infer a present pain solely from its behavioral effects; I undergo it.
3. **The Russellian proposal.** Phenomenal or protophenomenal character may supply the concrete nature that realizes physical structure, or at least our one positive clue to it. This remains an *abductive* step—an inference to the best explanation rather than a deduction.

The first two stages have considerable force. The third remains abductive. Silence in physics leaves theoretical room; it does not select the occupant.

### The Main Versions

**Russellian panpsychism.** Fundamental physical properties have phenomenal intrinsic natures. Microexperience plays the causal roles physics describes and helps constitute macroexperience.

**Russellian panprotopsychism.** The fundamental grounds are not experiences but possess a nature specially suited to constitute experience under the right conditions.

**Constitutive panqualityism.** Fundamental entities instantiate qualities without thereby becoming microscopic subjects of experience. This avoids the subject-summing problem aimed at views with many microsubjects, although it still owes accounts of macroquality, awareness, and subject constitution.

**Russellian realization theory.** Physical properties such as mass name dispositional roles realized by distinct phenomenal or protophenomenal categorical properties.

**Russellian identity and powerful-quality theories.** *Dispositional* describes what a property can do; *categorical* concerns the concrete character of the property that has the power. These descriptions may concern one property rather than two layers. On John Heil's powerful-quality view, qualitative nature and causal power cannot come apart. An a priori version claims the connection follows from the nature disclosed. An a posteriori version accepts a necessary unity that cannot be derived from the concepts alone, much like type-B physicalism's stopping point.

**The book's local Russellian option.** This label names a constructive possibility developed here, not a standard historical member of the Russellian taxonomy. Acquaintance may disclose something about a concrete organism-level physical episode while leaving its fundamental realizers open. If the episode's categorical character merely names the concrete realization of an organized, world-involving relational profile, the option converges with the book's type-B identity. A necessarily co-instantiated categorical realizer can remain metaphysically distinct, but it needs independent support or explanatory work to gain an abductive advantage.

### The Combination and Inheritance Problems

The familiar combination problem asks how microexperience yields the unified experience of a person. Chalmers separates subject-combination, quality-combination, and structure-combination. Phenomenal bonding, fusion, cosmopsychism, panqualityism, and Tye's transfer proposal address different parts of that bill and should not receive one generic objection.

The deeper issue concerns **inheritance**. Moving from an organism-level manifest quality to a qualitative microphysical ground requires a further premise:

> The manifest character disclosed in experience cannot consist in relational organization at the level of the whole and must ultimately derive from fundamental categorical qualities.

Neither introspection, causal closure, nor the structural character of physics establishes that premise. Panqualityism can avoid microsubjects, and powerful qualities can avoid freely swappable or causally idle realizers, while the descent to a qualitative fundamental base remains unsupported.

### Acquaintance and Revelation

Goff argues that phenomenal concepts may reveal part or all of the nature of their referents. His stronger real-acquaintance view says that attending to pain can disclose what pain essentially involves, not merely attach a label to an otherwise hidden property.

A physicalist should grant the datum without granting the promotion. Acquaintance can reveal the **manifest profile** of an experience: how red gets presented, how pain locates and evaluates a bodily disturbance, how one phenomenal character resembles another. It does not by itself reveal whether that profile has an absolutely intrinsic nature, depends on relational organization, or inherits its character from fundamental qualitative properties.

Marcelino Botin makes the underdetermination explicit: introspection need not present phenomenal character as absolutely intrinsic. Christopher Devlin Brown replies that phenomenal concepts may reveal nonfundamental physical categorical properties. That reply strengthens local disclosure while leaving the move to fundamental qualitative ground unestablished.

### My View

Russell's structural warning stands. A complete description in the vocabulary of physics may fail to confer every mode of knowing concrete reality, and no description substitutes for undergoing an experience. The epistemic residue remains real.

The metaphysical conclusion does not follow. The book treats strong representationalism as an independently constrained inference to the best explanation. Begin with a state already counted as conscious and borne by an identified subject. Its phenomenal character consists in world- or body-presenting content under a mode, with determinacy, field structure, and attention. Content-fixing history, state-consciousness, and subject attribution remain separate questions.

The view carries a visible defeater. A stable phenomenal contrast after the complete content-under-mode profile has genuinely been fixed would reopen the Russellian realizer question. Its absence supports the leaner view without proving that no local realizer exists. Without independent support or another explanatory payoff, the local option receives no abductive advantage; fundamental Russellian monism still depends on the inheritance bridge.

### Semantic Externalism

Semantic or content externalism holds that the contents of at least some thoughts depend partly on relations beyond the thinker's body. Intrinsically identical thinkers can therefore differ in what they believe or what their words refer to. The position concerns the individuation of content; it does not by itself claim that thoughts' physical vehicles extend outside the brain.

Environmental and social versions emerged prominently through Hilary Putnam's Twin Earth argument and Tyler Burge's arthritis case. Both challenge the idea that internal organization alone fixes every aspect of meaning, but they do not establish that all mental states have exclusively wide content.

### Putnam's Twin Earth

Imagine Earth and a counterpart planet whose lakes contain XYZ, a liquid with water's familiar observable properties but a different chemical constitution. In Putnam's premodern setting, speakers cannot distinguish the liquids chemically. Earth speakers use "water" for $H_2O$; their twins use the same sounds for XYZ.

Putnam argues that matching internal descriptions and dispositions need not give the two words the same extension. Their use belongs to different histories of interaction with the samples in their environments. The scenario idealizes the relevant intrinsic similarity; it is not a claim about the biological feasibility of replacing every molecule of water in a human body.

A related argument concerns the **division of linguistic labor**. People can refer to elm trees without being able to distinguish elms from beeches because they participate in a practice in which others possess that expertise. Reference need not depend on an identifying description known individually by every speaker.

Critics dispute whether these cases show differences in thought content, differences in linguistic reference alone, or a combination of wide and narrow aspects of meaning. The move from natural-kind terms to all concepts also needs separate support.

### Burge's Social Externalism

Burge considers someone who mistakenly believes that arthritis can occur in the thigh. In the actual community, the term concerns a disease of joints. In a counterfactual community, an otherwise similar term covers a broader range of rheumatic ailments. Holding the individual's relevant intrinsic states fixed, the two beliefs involve different socially determined concepts.

The argument concerns the person's partial understanding and deference to community usage. It does not require that the counterpart's thigh complaint actually be a rheumatic illness: changing the concept changes content, not automatically the truth of every corresponding belief.

Opponents question whether the original speaker really possesses the community's concept, rather than a mistaken private one. Burge's reply emphasizes ordinary practice: incomplete understanding does not normally prevent someone from thinking about the very condition on which an expert can correct them.

### Antecedents and Limits

Wittgenstein's discussion of private language and rule-following provides a related concern about standards of correct use. Its precise interpretation remains disputed. It should not be treated as a simple proof of Putnam's thesis, or as showing that thought always requires a presently existing linguistic community. A criticism of an essentially private correctness standard differs from a theory individuating concepts across environments.

Content externalism also differs from direct realism about perception. Environmental determination of a thought's content does not settle whether perceptual experience relates directly to objects, represents them, or involves sense-data.

### Narrow Content and Psychological Explanation

Internalists often distinguish **wide content**, such as reference to $H_2O$ rather than XYZ, from **narrow content**, shared by intrinsic duplicates. Narrow content might involve a conceptual role or a function mapping contexts to wide contents.

The motivation is explanatory: internally similar thinkers may make the same inferences and movements even when their wide contents differ. Externalists reply that world-involving contents support explanations of rational action, successful perception, and environmental error. Both sides must explain what their preferred content notion contributes and how it relates to neural mechanisms. Externalism challenges an entirely internal individuation of content, not every possible form of functionalism.

### Memory, Self-Knowledge, and AI

Slow-switching cases ask what happens when a person unknowingly moves between linguistic environments. Do old memories retain their original content, acquire new content, or permit several interpretations? Michael Tye argues that current environments can affect some memory contents; other externalists emphasize their original historical anchoring. The shift is disputed, not an automatic consequence of externalism.

A second problem concerns self-knowledge. How can someone know what they think without first investigating the environment that helps determine that thought? Compatibility theories distinguish knowing one's present thought from knowing all the conditions that individuate it. Critics press cases of unnoticed switching and apparent deductions about the external world.

For AI, the relevant environment can include linguistic data and social transmission as well as sensors and action. Fintan Mallory's teleosemantic treatment of word embeddings proposes content about word-types and their contexts. Whether that supplies particular-object reference or whole-system understanding remains a further question. Externalism alone does not deliver a verdict about every language model.

**Further Reading**

- Putnam, Hilary. 1975. "The Meaning of 'Meaning'." In *Philosophical Papers*, vol. 2, *Mind, Language, and Reality*. Cambridge University Press.
- Burge, Tyler. 1979. "Individualism and the Mental." *Midwest Studies in Philosophy* 4: 73–121.
- Tye, Michael. 1998. "Externalism and Memory." *Proceedings of the Aristotelian Society, Supplementary Volumes* 72: 77–94.
- Mallory, Fintan. 2026. "Teleosemantics for Neural Word Embeddings." *Mind & Language*. Published online August 9, 2026. doi:10.1111/mila.70037.
- Rowlands, Mark, Joe Lau, and Max Deutsch. 2020. "Externalism About the Mind." *Stanford Encyclopedia of Philosophy*.

**Related Writing on This Site**

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind, matter, and meaning.

**My View**

Semantic externalism does three jobs in my view:

**1. Grounding the Embodiment Argument**

If meaning doesn't supervene on internal physical states, then a system with the right current organization but a different history can carry different content. An always-envatted brain may refer to features of its feed, virtual kinds, or nothing determinate where the relevant history fails to select among rivals; it does not thereby inherit reference to the embodied objects its signals merely resemble. Putnam's externalism makes this point about language; my view extends it to perceptual and phenomenal content.

**2. Supporting the AI Critique**

An LLM's training is more than a timeless distribution. The corpus descends from speakers whose uses of *water* were anchored in public practice, and training selects internal mechanisms for responding to that record. Formal similarity among token patterns does not by itself fix $H_2O$ rather than water-contexts, but the corpus-mediated causal chain may carry some public reference into the mechanism when producer-consumer history preserves the relevant dependence. Mallory's word2vec analysis also gives a serious account of content about actual word-types and their linguistic distributions. These are worldly targets, even when they concern language. Whether that content is original at this limited grain remains unsettled; the inherited vocabulary, task, and loss do not by themselves make it derived. Originality requires history and producer-consumer use to fix accuracy conditions not exhausted by an interpreting practice, not freedom from semantic ancestry.

That grant does not settle whole-system understanding. The further question is whether content-bearing states enter a unified, persistent, world-sensitive economy of memory, correction, planning, and action. Nor does a stake or proprietor answer the semantic question: stake bears on autonomy and welfare, not on whether a mechanism has content.

### 3. Blocking Internalist Functionalism

The functionalist assumes mental content supervenes on functional role, which supervenes on internal physical organization. Semantic externalism breaks the first supervenience claim: same functional role, different content. This forces the functionalist either to retreat to narrow content (a technical notion that abstracts away from reference) or to abandon the claim that psychology is individualistic. My view takes the second option: minds are individuated partly by their environments, because content is.

### Simulation and Realization

A simulation models a target system; a realization instantiates a property or organization. The two categories can overlap. A weather simulation does not produce the modeled rainfall, but a simulated calculator can genuinely calculate. A controller implemented in software can genuinely regulate a boiler if its outputs affect that boiler. Calling an implementation a simulation therefore does not decide which properties it possesses.

The distinction enters philosophy of mind when researchers ask whether modeling a brain could also produce thought or consciousness. A faithful model establishes some correspondence with its target. Whether that correspondence preserves the properties relevant to mentality depends on a further theory of those properties.

### What Must Be Preserved?

Different tasks require different kinds of fidelity. A model may reproduce numerical outputs, imitate observable behavior, or preserve a network of causal dependencies. These achievements do not automatically coincide. A recording of a conversation reproduces its sounds without sustaining a conversation; a program that responds appropriately to new questions has a different causal capacity.

John Searle's hurricane and digestion comparisons emphasize that formal modeling need not reproduce every physical power of its target. Critics of that analogy answer that calculating and information processing differ from getting wet: the relevant property may concern organization rather than a particular material activity. The analogies locate the disagreement but cannot determine which category consciousness belongs to.

**Computational functionalists** hold that the right implemented causal-computational organization suffices for relevant mental properties. Some biological or substrate-sensitive theories require additional physical properties. Embodied and enactive theories emphasize interactions, historical development, or organismic organization. These positions offer different accounts of what a successful realization must preserve; none can establish its answer simply by repeating the word *simulation*.

### Genuine Computational Implementation

Implementation must constrain a physical system, not merely describe an actual sequence of events after it occurs. Triviality arguments associated with Hilary Putnam challenge accounts that allow arbitrary mappings between physical states and computational states. With sufficiently permissive mappings, ordinary objects appear to implement far too many programs.

David Chalmers replies that implementation requires an appropriate causal organization, including relationships among components and support for the relevant counterfactual transitions. If an input had differed, the corresponding physical transition should have differed in the way specified by the computation. A match to one recorded trajectory is insufficient.

This distinguishes running a controller from assigning controller labels to a wall. Disputes remain about which counterfactuals, state decompositions, and physical mechanisms count. Establishing genuine implementation also leaves open whether the implemented computation suffices for semantic content or consciousness.

**Content, Competence, and Consciousness**

Several questions often become compressed into "Does the simulation have a mind?" They can instead be examined separately:

- Does it implement the proposed computation or causal organization?
- Do its states represent anything, and what fixes their accuracy conditions?
- Does it use those states competently in a domain or integrate them into broader understanding?
- Is it conscious, and what supports that attribution?
- Which system or continuing subject bears the capacities?
- Does it have interests, vulnerability, or a basis for welfare?

This list is a way of organizing inquiry, not a consensus hierarchy or a theory that defines one achievement as necessarily preceding all the others. A subpersonal mechanism might represent without itself being a subject. A system might competently control a process without settling whether it understands its activity or experiences anything.

**Self-Maintenance and Practical Interests**

Enactive approaches connect autonomy and sense-making with maintaining an organization under precarious conditions. Ezequiel Di Paolo distinguishes autopoiesis, or self-producing organization, from **adaptivity**: regulation of activity relative to conditions of viability. A process that merely continues, such as a burning candle, need not monitor and alter its own activity in response to threats to its organization.

This motivates proposals that successful and unsuccessful trajectories acquire significance for a self-maintaining system. It does not establish a categorical moral obligation, nor does every theory of consciousness require metabolism or self-maintenance. Whether organizational norms suffice for semantic content, genuine interests, or welfare remains disputed. Functionalists may locate the relevant capacities elsewhere; critics of naturalized normativity ask how causal persistence can distinguish error from an observer's preferred outcome.

Artificial systems could satisfy some proposed regulatory conditions without being carbon-based. Conversely, representing an energy budget or a health variable does not by itself establish that the represented condition controls the persistence of the representing machinery. The actual dependencies need examination.

### System Boundaries and Platforms

A software agent depends on physical infrastructure. Some components may participate in its control and repair loops; others may be maintained independently by technicians or another system. A boundary drawn around the software, its hardware, or a distributed network can change the explanation being evaluated.

Reciprocal maintenance offers one proposed boundary criterion, while integrated information, causal control, cognitive accessibility, and biological individuality supply others. Nested systems make the disagreement concrete: a component, a robot, a colony, and a factory can each sustain an organized process without thereby constituting four equivalent conscious subjects. Corporate ownership and the shape of a casing do not settle these theoretical questions.

### Copies, Backups, and Virtual Environments

Copyability alone neither establishes nor excludes consciousness. A backup raises questions about continuity, duplication, and personal identity. It does not show that the running process lacks all relevant causal powers. Hardware-specific computation likewise does not automatically confer autonomy or experience.

Virtual objects are implemented through physical processes and can enter real causal relations with agents. A designer-built environment therefore need not be causally inert. The question is whether the agent's learning, action, and regulation depend on the virtual conditions in the ways a proposed theory requires. A numerical health score can either be a decorative label or govern the agent's continuing operation; those cases differ physically, not just verbally.

Open questions concern which organization consciousness requires, how content and understanding depend on history and interaction, and how to identify a subject in nested or distributed implementations. A general claim that every present AI system has or lacks a practical stake would require evidence beyond the distinction itself.

### Further Reading


- Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3: 417–424.
- Chalmers, David J. 1996. "Does a Rock Implement Every Finite-State Automaton?" *Synthese* 108: 309–333.
- Di Paolo, Ezequiel A. 2005. "Autopoiesis, Adaptivity, Teleology, Agency." *Phenomenology and the Cognitive Sciences* 4: 429–452.


### Related Writing on This Site

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind, matter, and meaning.

### My View

The slogan *simulation is not realization* survives only as a conditional:

> An implementation that merely represents a mind-making power does not realize that power. An implementation that genuinely possesses

> the required causal organization is not disqualified by being called a simulation.

For mind and stake, the relevant questions form a map rather than a single test:

1. Does the implementation possess observer-independent causal organization?
2. Did selection or training give any mechanisms content, and does that content remain derived or count as original?
3. Do several content-guided capacities support partial understanding, and how broadly and persistently do they integrate?
4. Which relation marks state consciousness, which persisting organization bears the state, and what world-presenting profile determines its character?
5. Does reciprocal self-maintenance give the organization autonomy, practical interests, and welfare?

Reasons-responsive agency cuts across questions 2–4 when represented reasons govern the bearer's conduct; it is not another grade of understanding or a proxy for consciousness or stake.

Different systems can occupy different positions on the map. Intrinsic computational functionalism may describe the entire package. If it can, I accept the description and inspect the realization. The dispute belongs in the architecture, not in the noun.

**The Symbol Grounding Problem**

The symbol grounding problem asks how a system's symbols acquire significance for that system, rather than only for an outside interpreter. Stevan Harnad's formulation challenges the idea that meaning can arise solely through formal relations among symbols. Consulting an unfamiliar monolingual dictionary illustrates the difficulty: every unknown word receives a definition made from other unknown words, with no independent way to identify what any of them concern.

The problem bears on symbolic theories of cognition and artificial intelligence. It does not begin with the assumption that every computer lacks meaning. It asks what would make an attribution of meaning explanatory rather than merely imposed.

**Harnad's Proposal**

Harnad connects basic symbols with nonsymbolic capacities to discriminate and categorize sensory input. A system can acquire a category through interaction and corrective feedback, learning which objects or events count as instances. Symbols for those categories can then enter compositions that express more complex classifications and claims.

His later distinction between **sensorimotor toil** and **symbolic theft** describes two routes to category acquisition. Trial and error can establish a category directly; linguistic descriptions can communicate a new category using symbols already grounded. A child need not encounter every dangerous animal personally to learn about it. The linguistic route nevertheless presupposes enough established category knowledge to make the description useful.

A robotic system that sorts red objects might connect a symbol with a learned sensory category. But successful sorting does not immediately answer whether the robot understands a word as a

human does, grasps its broader relations, or has a conscious experience of red. Harnad explicitly distinguishes a theory of grounding from a complete theory of meaning.

His own additional condition concerns **feeling**: in the 2002 account, a thought has full meaning only when it is grounded and there is something it feels like for the thinker to think it. Grounding supplies an autonomous connection between symbols and categories; it does not, on his account, establish felt thinking. This consciousness condition belongs to Harnad's theory, not to every account of meaning. Externalist, teleosemantic, and functionalist approaches need not impose the same condition, and it does not by itself decide whether any particular robot has experience.

### Connections with the Chinese Room

John Searle's Chinese Room imagines someone manipulating unfamiliar symbols according to formal rules while producing appropriate responses. Searle argues that implementing a program alone does not suffice for understanding. Harnad asks what nonsymbolic resources could connect the symbols with their categories.

The arguments overlap without being identical. A sensorimotor solution can meet a grounding demand while leaving Searle's concern about understanding unresolved. The systems reply also challenges the choice of subject: the person following the rules might not understand, while the larger organized system could. A grounding theory must therefore state both what makes its representations meaningful and to which system it attributes the relevant capacity.

### Social, Historical, and Distributional Routes

Sensorimotor interaction is one proposed source of grounding. Externalist theories emphasize causal history and participation in public linguistic practices. Teleosemantic theories appeal to the selected functions of mechanisms that produce and use representations. These approaches may complement sensory categorization or contest the claim that every meaningful symbol must receive its grounding through the individual's own sensory contact.

Testimony provides a familiar case. Someone can refer to a distant planet without observing it directly. The connection may run through other speakers, instruments, and an inherited naming practice. Critics ask whether transmitted reference supplies the system's own semantic competence or merely the possibility of interpreting its tokens.

Fintan Mallory's analysis of word2vec adds a distinct target: an embedding can represent facts about a word-type's distribution in linguistic contexts. On his consumer-based account, those linguistic facts belong to the world, even if the embedding does not independently represent the tree or person named by the word. Content about *the word "tree"* differs from reference to a particular tree.

### Language Models and Further Questions

Training on text supplies a history of interaction with linguistic artifacts, not an encounter with a causally disconnected mathematical object. Nevertheless, successful prediction does not by itself tell us which content a mechanism carries or whether the system understands a conversation. Different theories assign different significance to training objectives, consumers of internal states, public linguistic ancestry, and subsequent correction.

Multimodal training, tools, sensors, persistent memory, and action can change those relations. None serves as a universal semantic switch. An image classifier can associate a visual pattern with a label while remaining limited in how it uses that label; a more integrated agent may deploy similar information in inference and planning.

Strong objections run in both directions. Grounding advocates ask how a purely formal network distinguishes truth from useful association. Their opponents ask why well-organized inferential and social use could not already constitute meaning, and whether sensorimotor success explains semantics or only reliable behavior. Even if grounding were settled, consciousness and the scope of understanding would require further argument.

**Further Reading**

- Harnad, Stevan. 2002. "Symbol Grounding and the Origin of Language." In *Computationalism: New Directions*, edited by Matthias Scheutz, 143–158. MIT Press.
- Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3: 417–424.
- Mallory, Fintan. 2026. "Teleosemantics for Neural Word Embeddings." *Mind & Language*. doi:10.1111/mila.70037.

**Related Writing on This Site**

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind, matter, and meaning.

**My View**

Form alone cannot fix distal reference, but Harnad-style sensorimotor grounding is one route to world-directed content rather than the only route. Public practice and corpus-mediated causal history can transmit derived reference; selection, learning, and world-involving use can establish original content where they fix accuracy conditions not exhausted by an interpreting practice. Mallory gives a serious account of content at the linguistic-distribution grain. Whether that content is original remains unsettled: inherited vocabulary and training design alone do not decide the issue. Failing to refer to a particular tree does not refute content about the word *tree*. None of this by itself establishes whole-system understanding or inner life. The symbol grounding problem shows why syntax alone cannot manufacture semantics, not why every current model, service, or agent lacks all semantic content.

**Teleosemantics**

Teleosemantics explains representational content through function. A state represents a condition partly because a mechanism has the function of producing or using that state in relation to that condition. Many versions ground function in a history of selection, reproduction, learning, or differential retention. The aim is to explain meaning and misrepresentation without already assuming a mind that interprets the state.

Error supplies a central motivation. A signal can remain causally effective when what it represents is absent. Mere correlation does not yet distinguish a correct signal from an error: a mechanism can correlate perfectly with one proximal feature while misleading its consumer about a different condition.

### The Frog and the Disjunction Problem

A frog's visual system responds to small dark moving objects and triggers tongue movements. What does the signal represent: a fly, edible prey, or a small dark moving object? All may have coincided in the history that shaped the mechanism. A BB pellet separates the candidates.

Teleosemantics asks which condition explains the mechanism's successful functioning and retention. On a consumer-oriented reading, obtaining food can privilege prey over arbitrary dark objects. On a producer-oriented or causal-informational reading, the visual mechanism's selected sensitivity helps constrain content. Neither label automatically fixes the exact answer.

The problem is **functional indeterminacy**: several descriptions can explain the same selected effect. If the signal represents small dark moving things, responding to a BB may not be a representational error, even if it harms the frog. If it represents edible prey, the same response can involve misrepresentation. Defenders appeal to details of selection, producer-consumer organization, and systematic mappings; critics argue that those details may still leave multiple contents eligible.

### Major Approaches

**Ruth Millikan's consumer-based account** places representations within historically established producer-consumer systems. Consumers' proper functions help determine the conditions under which representations enable successful performance. "Normal" functioning in this framework concerns an explanatory history, not simply what happens most frequently.

**Fred Dretske's informational approach** distinguishes a state's actually indicating a condition from its having the function of indicating it. Learning can recruit an indicator for control of behavior. Once it has that function, the state can occur when the condition is absent and so misrepresent.

**David Papineau and Karen Neander** develop different accounts of how selected effects, successful action, and causal-informational relations constrain content. The family therefore does not have one agreed mechanism or one uniform answer to the frog case.

Some representations combine descriptive and directive roles. Millikan's *pushmi-pullyu* representations both indicate a condition and guide a response, as a bee's dance can indicate a food location while directing other bees there. These roles need not be separately represented in a simple system.

### History and Swampman

The Swampman thought experiment imagines a molecule-for-molecule duplicate produced without the history of the person it duplicates. Strict historical versions of teleosemantics deny the new duplicate the content its counterpart possesses at the instant of creation, despite matching current behavior and organization.

Critics regard that verdict as evidence that history cannot be necessary for every representation. Defenders answer that the example removes exactly the explanatory relation on which content depends and that later interaction or learning may supply relevant history. Other theories soften

historical conditions or appeal to present organization. Self-maintenance alone does not restore the missing selection history on a strictly etiological theory.

### Neural Networks and Levels of Attribution

Fintan Mallory applies consumer-based teleosemantics to word2vec. He treats training as a history establishing producer-consumer functions and argues that embeddings combine content about word-types' contextual distributions with directions to downstream mechanisms. The numerical encoding alone does not provide the content; the learned organization and its use do explanatory work.

This proposal challenges an all-or-nothing choice between representing ordinary objects and representing nothing. It also raises objections: does optimization constitute the relevant selection process, do human-defined tasks make the content merely derived, and how exactly do actual linguistic distributions constrain the internal mapping? Mallory defends genuine content at this limited linguistic grain. That does not establish a model's understanding of everything named in its training texts.

Content attribution to a mechanism differs from attributing belief, perception, or understanding to a whole agent. Teleosemantics need not posit a separate owner that activates each representation, but identifying which system uses the content in a cognitive capacity remains a further explanatory task. Biological autonomy and welfare theories pose related questions without supplying a universally accepted extra semantic condition.

### Enactivist Objections and Open Problems

Radical enactivists such as Daniel Hutto and Erik Myin question whether natural functions and covariance produce truth-evaluable content at all. They allow purposeful interaction while resisting its automatic description as representation. Teleosemanticists reply that historical functions can establish satisfaction conditions and distinguish malfunction from mere unusual behavior. The objection is whether that explanation earns semantic vocabulary, not whether the mechanisms possess causal organization.

Other challenges concern novel and abstract concepts, the distinction between malfunction and representational error, and how to move from simple biological signals to articulated human thought. Teleosemantics remains an influential family of naturalizing theories, not a settled solution to every kind of meaning.

### Further Reading


- Schulte, Peter, and Karen Neander. "Teleological Theories of Mental Content." *Stanford Encyclopedia of Philosophy*, substantive revision 2022.
- Mallory, Fintan. 2026. "Teleosemantics for Neural Word Embeddings." *Mind & Language*. doi:10.1111/mila.70037.


### Related Writing on This Site

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind, matter, and meaning.

### My View

A family of states, a correspondence rule, and a producer-consumer history can establish content at the grain the explanatory organization supports. It counts as original to the extent that history and use fix accuracy conditions not exhausted by an interpreting practice. The mapping must explain stabilization and survive relevant interventions. Mallory's account of content about actual word-types and distributions deserves that test; originality at this limited grain remains unsettled. The vocabulary, corpus, task, and loss have a human ancestry, but ancestry alone does not show that their interpreting practice exhausts the mechanism's accuracy conditions. A coordinated reparameterization can preserve the word-to-context relation; changing the external pairing changes a candidate content-fixing relation. Neither maneuver alone refutes contextual content or settles its independence. No further proprietor relation turns content on.

Content at one mechanism remains compatible with ignorance at the level of the whole. The book's later account must earn perception, belief, understanding, and phenomenal attribution through the organization and poising of content-bearing states rather than through ownership as a semantic ingredient.

### The Church-Turing-Deutsch Principle

The Church-Turing-Deutsch principle concerns the possibility of a universal physical simulator. David Deutsch formulated a physical Church-Turing principle in 1985: "every finitely realizable physical system can be perfectly simulated by a universal model computing machine operating by finite means." His paper connects this proposal with the universal quantum computer.

This differs from the ordinary Church-Turing thesis, which identifies effectively calculable functions with functions computable by a Turing machine. The physical principle concerns what machines and physical laws permit. It is not a theorem that automatically follows from the mathematical theory of computability.

### From Calculability to Physical Simulation

The ordinary thesis connects an informal idea of an effective procedure with precise mathematical models. Equivalence results among standard models support that identification, but they do not prove that every physically possible process obeys the thesis.

Deutsch asks which physical theory supports universal simulation. Classical models allowing unrestricted continuous quantities create difficulties for exact simulation by finite means. Quantum theory motivates a different universal machine, whose operations are themselves physical and quantum. His proposal seeks to locate computational universality within physics rather than treat physics as an incidental implementation of abstract algorithms.

Deutsch defines **perfect simulation** through computational equivalence under specified labellings of the target's and simulator's inputs and outputs. A program must make the simulator functionally indistinguishable from the target in the corresponding experiments. For probabilistic outputs, the requirement concerns matching probability distributions, not matching each individual random outcome. This operational equivalence does not require duplicating every physical property of the target.

That defined task should be distinguished from approximating an evolution or reproducing selected observations without establishing computational equivalence. Claims about universality must specify the systems, observables, resources, and accuracy at issue. Idealized universality does not promise a practical machine capable of handling every target with available time and memory.

**Quantum Computability and Efficiency**

Quantum computation exploits superposition, interference, and other quantum operations. It can change the resources required for some computational tasks and provides a natural route to simulating quantum systems.

This does not imply that an ordinary universal quantum computer computes nonrecursive functions. Computability asks whether a function can be computed at all within a model; complexity asks how many resources its computation requires. A quantum speed advantage addresses the latter question without automatically overturning the former. Conversely, mathematical computability does not guarantee an efficient classical simulation of a physical quantum process.

Thus the labels *physical Church-Turing thesis*, *efficient simulation thesis*, and *Deutsch's principle* should not be used interchangeably. They make claims of different strength.

**Computationalism and Philosophy of Mind**

Gualtiero Piccinini distinguishes the claim that neural activity can be computationally modeled from the empirical claim that neural mechanisms actually perform computations explaining mental capacities. The ordinary Church-Turing thesis applies once effective calculability has been established; it does not establish that a brain is a computing mechanism. On his analysis, even a broad physical simulability claim would not by itself identify the specific mechanism explaining cognition. This is a criticism of arguments for computationalism, not a proof that computationalism is false.

The physical principle bears on computationalism through a further argument: if a brain can be simulated, and the simulation preserves every organization sufficient for mentality, then the implementation may possess the corresponding mental capacities.

The second premise needs a theory of mind. A computational functionalist argues that the relevant causal organization can be realized in different substrates. A biological or substrate-sensitive theorist may require powers not preserved by the chosen model. Content externalists may ask which relations to an environment or history a proposed duplication preserves. Deutsch's principle alone does not adjudicate these accounts.

A weather model's failure to make its surroundings wet illustrates that not every modeled property transfers. It does not prove that thinking resembles rainfall rather than calculation. A simulated calculator can calculate, and a simulated controller can control a real process. Both proponents and critics must identify the property at issue rather than infer the answer from the simulation label.

Nor does an arbitrary interpretation establish genuine implementation. The causal relations supporting the model's counterfactual transitions remain relevant. Faithful outputs from one run differ from realizing a system capable of responding appropriately under alternative inputs.

**Timpson's Criticism**

Christopher Timpson distinguishes the mathematical Church-Turing hypothesis from a physical Turing principle about universal simulation. He argues that the physical principle does not supply the foundation for effective calculability that Deutsch assigns to it. A claim about what physical systems can simulate cannot simply replace an analysis of effective mathematical procedures.

This objection concerns the relationship between two theses. It need not deny the usefulness of quantum computing or refute every universality result. Deutsch's approach motivates treating computation as a physical process; Timpson's criticism insists that physical simulability and mathematical calculability retain distinct roles.

**Unresolved Scope**

Questions remain about exact versus approximate simulation, finite resources, the treatment of continuous quantities, and the physical assumptions of universality. In philosophy of mind, additional questions concern content, understanding, consciousness, and the identity of the system bearing them.

Accepting universal physical simulation therefore need not entail computational functionalism. Rejecting that entailment likewise does not establish that silicon or virtual systems cannot think. The principle supplies a proposal about simulation; arguments about mental realization need further premises.

**Further Reading**

- Deutsch, David. 1985. "Quantum Theory, the Church-Turing Principle and the Universal Quantum Computer." *Proceedings of the Royal Society of London A* 400: 97–117.
- Piccinini, Gualtiero. 2007. "Computationalism, the Church-Turing Thesis, and the Church-Turing Fallacy." *Synthese* 154: 97–120.
- Timpson, Christopher G. 2004. "Quantum Computation and the Church-Turing Hypothesis." Chapter 6 of *Quantum Information Theory and the Foundations of Quantum Mechanics*, DPhil thesis, University of Oxford.
- Chalmers, David J. 1996. "Does a Rock Implement Every Finite-State Automaton?" *Synthese* 108: 309–333.

**Related Writing on This Site**

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind, matter, and meaning.

**My View**

I am glad to give Deutsch his insight, which is real and is his. The standard Church-Turing thesis *does* carry a hidden physical commitment, and dragging it into the light was a genuine contribution — one that, characteristically, also produced the universal quantum computer as a

by-product. Where physics is concerned I have no quarrel; I will even grant that the brain can be perfectly simulated by Deutsch's universal machine, quantum and all.

What I deny is that any of this rescues computational functionalism. CTD closes the first gap, from mathematical computability to physical simulability, but leaves the realization question open. Abstract dynamical form alone does not fix content. World-involving history may fix what a state represents; whole-system understanding additionally requires content-guided capacities integrated in one persisting, answerable cognitive economy; and consciousness further requires appropriately poised, perceptually organized content within an identified bearer. CTD supplies none of those verdicts.

Two clarifications, because the argument gets over-read in my favor. The word *simulation* carries no verdict of its own: a system built in silicon, or running as software, might realize that organization rather than merely depict it, and if it does, calling the thing a simulation settles nothing against it. And the standing question about text-only systems belongs to the evidence, not to the definition. Considered apart from persistent memory, tools, and world-coupling, a bare language model has not established integrated, world-involving understanding, and whether anything it does amounts to experience remains open — but I would not read either answer off its manufacture, its formalism, or the borrowed ancestry of its vocabulary. What CTD cannot do is decide any of this, because it speaks to form alone.

So I treat CTD as the computationalist's *best* card, and that is exactly why it is worth playing out in full. The card is real, the physics is sound, and the principle wins everything it claims — but what it claims is simulability, not a verdict about realization. Timpson presses the point harder, doubting that the principle even secures its physical footing. He may well be right. But the case does not depend on it. Grant Deutsch the whole of his principle, and the question whether a particular simulation realizes a mind remains unanswered.

**The Explanatory Gap**

The explanatory gap concerns our inability to explain why a physical process has a particular conscious character. Joseph Levine introduced the term in "Materialism and Qualia: The Explanatory Gap" in 1983. His argument does not conclude that physicalism is false. It asks what remains unexplained even if a proposed mind–brain identity is true.

A description of the neural processes involved in pain can explain causes, behavioral consequences, and interactions with other processes. It seems to leave a further question: why does that process feel like this, rather than like something else or nothing at all? The gap concerns explanation and knowledge; whether it also reveals a difference in the kinds of properties that exist remains disputed.

**Levine's Argument**

Levine contrasts familiar scientific identities with psychophysical identities. Identifying heat with molecular motion can become intelligible through an account of the causal role associated with heat. In the proposed identity "pain is C-fiber firing," the physical description appears not to make the felt character similarly intelligible.

C-fiber firing here serves as a schematic example, not an adequate contemporary neuroscience of every pain. The philosophical issue survives substitution of a more realistic neural description if the same explanatory question remains.

Levine recasts a metaphysical challenge associated with Saul Kripke into an epistemological one. Conceiving an identity as false does not, in Levine's argument, establish that it actually is false. It instead exposes an explanatory limitation and a difficulty determining which physical identity claims are correct. Treating his position as a straightforward endorsement of dualism misses this restraint; treating it as a settled vindication of physicalism misses the challenge.

**Gap, Identity, and Necessity**

An empirically discovered identity can be necessary even when its necessity cannot be read off the concepts involved. Physicalists use this distinction to argue that a conceptual gap need not entail two properties. Opponents respond that phenomenal concepts differ from ordinary scientific concepts: they seem to present what the property itself is like, making the residual ignorance especially difficult to explain.

Both sides can agree that an identity, if true, relates one thing under different descriptions. Their disagreement concerns whether the proposed identity explains its target and whether the epistemic evidence supports it. Declaring an identity necessary does not independently establish that it is true.

**The Phenomenal Concept Strategy**

One influential physicalist response locates the gap in how people think about experience. Phenomenal concepts apply to experience from a first-person standpoint; physical concepts characterize it through scientific descriptions. If these concepts differ in their cognitive roles while referring to the same physical property, their separation might explain why the identity remains opaque.

Versions appeal to recognition, demonstration, acquaintance, or the incorporation of an experienced instance into a concept. Brian Loar's recognitional account provides one important example. A person who sees red can acquire a capacity to recognize its appearance that a description alone does not confer.

Mary's Room illustrates the dispute. Mary knows the physical account of color vision before experiencing color. On a physicalist reading, her release supplies a new capacity or way of knowing an already known property. On an antiphysicalist reading, she discovers a fact absent from the complete physical account. Naming a new mode of access does not by itself decide between those readings.

Michael Tye's later *Consciousness Revisited: Materialism without Phenomenal Concepts* complicates the taxonomy: acquaintance can remain central to a physicalist response even when the theory rejects a special class of phenomenal concepts. The strategy is a family of proposals, not one universally adopted mechanism.

### Objections and Replies

David Chalmers asks whether explaining the distinctive features of phenomenal concepts can do all the required work in physical terms. If the account omits the features tied to actual experience, it may explain judgments without explaining phenomenal knowledge. If it includes those features, critics argue that it risks presupposing the consciousness it was meant to explain.

Physicalists answer that a conceptual account need not deduce experience from physics to block an inference from ignorance to nonphysical properties. They still owe a positive account of the relevant concepts and independent support for the proposed physical identity. Representationalist accounts offer such identities, but critics ask whether represented colors, bodily conditions, and modes of presentation exhaust the felt character.

### Relation to the Hard Problem

The hard problem asks why physical or functional processes involve experience at all. The explanatory gap focuses on why a proposed physical identity remains unintelligible. The questions overlap, but neither terminology dictates a complete metaphysical position.

Chalmers uses explanatory limitations to motivate nonreductive approaches. Type-B physicalists accept a persistent epistemic gap while denying an ontological gap. Other approaches expect explanatory progress, question the initial demand, or attribute the appearance of special phenomenal properties to cognitive mechanisms. The unresolved issue is what an adequate explanation would require and how much the present gap tells us about reality.

### Further Reading


- Levine, Joseph. 1983. "Materialism and Qualia: The Explanatory Gap." *Pacific Philosophical Quarterly* 64: 354–361.
- Loar, Brian. 1990. "Phenomenal States." *Philosophical Perspectives* 4: 81–108.
- Tye, Michael. 2009. *Consciousness Revisited: Materialism without Phenomenal Concepts*. MIT Press.
- Chalmers, David J. 2007. "Phenomenal Concepts and the Explanatory Gap." In *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, edited by Torin Alter and Sven Walter, 167–194. Oxford University Press. doi:10.1093/acprof:oso/9780195171655.003.0009.


### Related Writing on This Site

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind, matter, and meaning.

### My View

My deployment of the explanatory gap is threefold.

**Accept the diagnosis.** Yes, there is an epistemic gap between third-person physical description and first-person phenomenal acquaintance. Levine is right about this.

**Use acquaintance and recognition.** The gap is real but not ontologically loaded. Description and first-person recognition provide different routes to one physical fact; explaining the recognitional capacity does not confer it.

**Diagnose the residual mystery as inner-theater inheritance.** When the gap *feels* like more than an epistemic gap — when it feels like evidence for non-physical phenomenal properties — the inner-theater picture is doing that work in the background. Name the picture and drop it, and the residual feeling loses its force. The intuition that "something is still left out" trades on the picture; the picture is what gets left out, not a phenomenon.

I treat Levine's gap as a useful diagnostic tool, a way of marking exactly what physicalism does and does not make available *a priori*, without treating the gap as a second feature of nature. The acquaintance-recognition account explains why the gap persists given the identity; it cannot establish the identity. Chapters 6 and 11 carry those burdens separately.

### The Hard Problem of Consciousness

The hard problem asks why physical processing involves conscious experience. David Chalmers introduced the terminology in "Facing Up to the Problem of Consciousness" in 1995. Explaining how a system discriminates stimuli, remembers events, directs attention, or reports internal states appears to answer one set of questions. Why there is something it is like to undergo those processes appears to raise another.

The distinction remains contested. Some philosophers treat it as a genuine barrier to reductive explanation; others argue that a complete physical account would close the gap or explain why people believe one remains. Formulating the question does not by itself establish dualism or any particular theory of consciousness.

### Easy Problems and Experience

Chalmers calls problems about discrimination, integration, reportability, attention, behavioral control, and wakefulness "easy" because a mechanism performing the relevant function appears to supply the required explanation. He does not mean that the scientific work is simple or already completed.

The hard problem concerns phenomenal consciousness: the visual appearance of red, the sound of a clarinet, or the felt character of pain. A complete account of the corresponding discrimination and report mechanisms still seems to permit the question of why those mechanisms involve experience. Chalmers argues that explaining another function does not automatically answer it.

Critics challenge either the separation of functions from experience or the explanatory demand it imposes. A conscious activity may be characterized inadequately if treated as nothing beyond input-output behavior. Conversely, distinguishing a further target in words does not guarantee a further property in nature.

### The Explanatory Gap and Knowledge Arguments

Joseph Levine's explanatory gap asks why a true mind–brain identity would remain unexplained. His 1983 argument explicitly stops short of establishing physicalism's falsity. The hard problem

shares that explanatory concern but supports a wider debate about whether standard reductive methods can succeed.

Mary's Room makes a related challenge epistemic: would someone knowing every physical fact about color vision learn something upon first seeing color? Antiphysicalists take a new fact to reveal the incompleteness of physical knowledge. Physicalists propose new abilities, acquaintance, or a different concept of the same fact. Both the formulation of "every physical fact" and the classification of Mary's gain require argument.

**Zombies and Conceivability**

A philosophical zombie duplicates a conscious person physically while lacking experience. Chalmers argues that suitably idealized conceivability can support a claim of possibility, which would threaten physicalism. This is a modal argument, not a proposal to find zombies in a laboratory.

Physicalist replies differ. Some deny that the complete zombie scenario is ideally conceivable. Others grant an epistemic opening but deny the inference to metaphysical possibility, comparing the situation with empirically discovered necessary identities. Externalist theories additionally ask what environment and history the duplication includes. Matching an internal functional profile is weaker than duplicating every physical fact relevant to the theory under examination.

An argument that grants conceivability must defend its rejection of the possibility inference. An argument that rejects conceivability must explain why the scenario nevertheless appears coherent. Neither answer follows solely from confidence in a preferred theory.

**Nonreductive Approaches**

Chalmers proposes treating experience as fundamental and seeking principles relating it to physical organization. In the 1995 paper, these include structural coherence and organizational invariance. The latter allows the same fine-grained causal organization to support experience across different substrates; his rejection of reductive explanation therefore should not be confused with a rejection of artificial consciousness.

Russellian monist and panpsychist approaches seek consciousness-relevant natures within the basis of matter rather than adding experience to an entirely nonexperiential foundation. They face questions about combination, subject unity, and explanatory constraint. Interactionist and other dualisms encounter further problems about mental causation and their relationship with physical explanations.

**Physicalist and Deflationary Replies**

Type-A physicalists expect an adequate functional or physical explanation to remove the supposed conceptual residue. Type-B physicalists accept an epistemic gap while defending an a posteriori identity. Phenomenal-concept strategies try to explain why different ways of thinking can concern one physical property without making the identity transparent.

Representationalists identify phenomenal character with appropriately qualified representational content. They appeal to the world-directed character of experience and to contrasts in what or

how a state represents. Critics reply that representation may occur unconsciously and that phenomenal differences may persist despite matching content. Representationalists must specify the consciousness-relevant conditions and defend the claim that their account preserves experience.

Illusionist approaches explain the apparent possession of special phenomenal properties through introspective and cognitive mechanisms. Critics ask whether this changes the subject; defenders argue that explaining the appearance is precisely the task. Their disagreement cannot be settled by treating either side's description of the target as neutral.

**Open Questions**

Scientific theories investigate neural organization, access, recurrence, integration, and the distribution of consciousness. Their empirical success may constrain the philosophical alternatives without automatically resolving the explanatory dispute. The remaining questions concern what counts as a complete explanation, whether conceivability tracks possibility, and how first-person evidence bears on a physical theory.

**Further Reading**

- Chalmers, David J. 1995. "Facing Up to the Problem of Consciousness." *Journal of Consciousness Studies* 2: 200–219.
- Levine, Joseph. 1983. "Materialism and Qualia: The Explanatory Gap." *Pacific Philosophical Quarterly* 64: 354–361.
- Tye, Michael. 2009. *Consciousness Revisited: Materialism without Phenomenal Concepts*. MIT Press.

**Related Writing on This Site**

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind, matter, and meaning.

**My View**

My view takes the hard problem seriously. The gap is real, but epistemic: it lies between public description and an acquaintance-grounded recognitional way of reaching the same fact, not between two kinds of property in nature.

The inner-theater picture helps explain why the gap feels ontological. It casts phenomenal character as intrinsic mental paint added to physical and representational organization. Transparency presses against that picture, but it supplies evidence rather than a deduction. The identity claim rests on the cumulative case that, within a state already counted as conscious and borne by an identified subject, phenomenal character consists in the complete world-presenting intentional profile: independently fixed content under a perceptual or bodily-affective mode, with determinacy, field structure, and attentional organization.

Content fixing does not by itself establish consciousness. State consciousness remains a further empirical question about poising, recurrence, global availability, or related integration; subject individuation asks which persisting control organization bears the state. If the identity claim holds, no extra phenomenal property waits to be connected to that organization. The hard problem

is then deflated rather than solved: the ontological bridge disappears, while the conceptual gap and the questions of realization and distribution remain.

### The Inner Theater

The inner theater is a critical label for a picture of conscious experience as an internal display presented to an observer. It brings together metaphors of a mental screen, images inside the head, and a spectator who encounters them. Daniel Dennett's more specific "Cartesian Theater" targets the supposed place or stage in processing where information becomes definitively conscious.

Few philosophers endorse a literal theater. The label belongs mainly to critics, and one task is to determine whether a particular theory actually depends on the picture. Believing in mental representations, phenomenal qualities, or first-person knowledge does not automatically commit someone to an internal spectator.

### Historical Background

Early modern theories gave ideas a central role in explaining perception and knowledge. Cartesian mind–body dualism also distinguished thinking substance from extended matter. Later critics connect these traditions with an image of a subject inspecting mental items and inferring an external world from them.

That reconstruction compresses important differences among philosophers. Descartes's account of sensation, Locke's theory of ideas, and later sense-datum theories are not one doctrine. Nor does acknowledging an experienced appearance always imply that a person first inspects a private object and then consciously reasons toward the world. The metaphor can identify a structural worry without supplying an adequate history by itself.

### Several Distinct Commitments

The theater picture can involve different claims:

1. Representations occupy an inner space that functions like a display.
2. Consciousness consists in observing items on that display.
3. Perceiving an external object requires first perceiving an internal intermediary.
4. A central point marks the transition from unconscious processing to conscious presentation.
5. First-person access to the supposed items is infallible.

These claims need not travel together. A theory might posit mental intermediaries without infallibility, or a central cognitive workspace without an observer who watches it. Distinguishing the commitments prevents a criticism of one claim from being treated as a refutation of every theory of inner states.

### The Homunculus Problem

An inner observer cannot explain perception if its own perceptual capacities require the same explanation again. If the brain projects an image for a little spectator, how does that spectator see it? Repeating the display-observer arrangement generates a regress.

Functional explanations can sometimes use apparently agent-like components without this regress. A detector, comparator, or selector may perform a restricted task that can itself be

explained mechanically. The problem arises when a component receives the unexplained capacity of the whole subject, not whenever a scientist describes a subsystem as detecting information.

### Dennett's Cartesian Theater

In *Consciousness Explained*, Dennett rejects a central finish line at which distributed processing becomes a definitive presentation to consciousness. His target includes "Cartesian materialism": rejecting an immaterial soul while retaining a privileged material theater where everything comes together.

Temporal perception supplies a concrete difficulty. Different signals reach different brain locations in different orders. Selecting one arrival point as the order of experience requires an explanation of why that location has the proposed authority. Dennett's Multiple Drafts model instead emphasizes distributed processing and the subsequent influence of discriminations, without one final internal presentation.

Critics argue that dispensing with a definitive theater may leave the unity or felt character of experience insufficiently explained. Dennett disputes the conception of phenomenal properties driving that objection. It would be inaccurate to describe his position simply as deleting all experience; what experience amounts to is part of the disagreement.

### Transparency, Representation, and Knowledge

Transparency arguments observe that attention to visual experience ordinarily returns to the presented colors, shapes, and scene. Michael Tye takes this to support a representational account: an experience need not display its neural vehicle or private colors to an inner observer. Tim Crane questions how much introspection can establish about the underlying nature of experience.

The vehicle/content distinction makes a related point. Neural processes can be located inside a head while representing an external apple. The location of the vehicle does not establish that the perceived apple is an inner object. This permits representation without a literal screen.

Sellars's criticism of the Myth of the Given challenges another commitment: allegedly immediate sensory items cannot automatically justify beliefs independently of the conceptual capacities needed to treat them as reasons. This epistemological criticism differs from an argument against neural imagery or every form of perceptual consciousness.

### Imagery and Remaining Disputes

Closing one's eyes and imagining an apple can invite the theater metaphor. A theory must explain the experience's visual character, its relation to perception, and its use in memory and planning. Rejecting a spectator does not make those phenomena disappear.

Sense-datum theorists can deny a homunculus while defending direct awareness of sensory particulars. Qualia realists can defend properties of experience without positing a second observer. Workspace theories can propose integration and broadcast without a mental screen. Each account should be assessed through its actual explanatory commitments.

The open questions concern the unity of experience, the organization of imagery, the limits of introspection, and whether any central structure has a privileged role in consciousness. Anti-theater arguments constrain answers; they do not by themselves solve every problem of mind.

**Further Reading**

- Dennett, Daniel C. 1991. *Consciousness Explained*. Little, Brown.
- Crane, Tim. 2000. "Introspection, Intentionality, and the Transparency of Experience." *Philosophical Topics* 28: 49–67.
- Tye, Michael. 1995. *Ten Problems of Consciousness*. MIT Press.

**Related Writing on This Site**

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind, matter, and meaning.

**My View**

The inner theater is the central misleading picture in philosophy of mind, and dismantling it is at the core of my project. Naming the picture matters. As long as it operates implicitly, it generates problems that look like discoveries about the mind; named and rejected, it becomes recognizable as a misleading metaphor to avoid.

**The Self**

"The self" can name a subject of experience, a continuing person, an embodied agent, or an autobiographical identity. These are connected questions, but they are not interchangeable. A theory of how someone tells a life story need not identify what experiences the story; a theory of present subjectivity need not establish what makes the same person persist for decades.

Philosophical accounts range from enduring egos and bundles of mental events to bodily, psychological, phenomenological, and narrative theories. Their disagreements concern both what exists and which explanatory task a conception of self should perform.

**Ego and Bundle Theories**

The Cartesian ego theory treats a person as a continuing thinking substance. Different experiences belong to one subject because that substance undergoes them. Its attraction is a direct account of ownership and persistence that does not identify a person with any particular passing experience.

A bundle theory, associated with David Hume, denies that a separate ego must stand behind the connected perceptions. Hume's introspective challenge asks whether examination reveals such an ego rather than particular perceptions. Michael Tye's discussion distinguishes the continuing substance from an organized, temporally extended cluster of mental events.

The introspective argument has limits. Failure to encounter a self as another object does not alone establish that there is no subject. Ego theorists can deny that subjects should appear in the same way as their experiences. Bundle theorists, meanwhile, must explain which relations unite the events into one person's life rather than an arbitrary collection.

### Unity and the Binding Question

Several perceptions can occur together within one conscious field. One person can also undergo changes of attention, memory, mood, and bodily condition over time. An account of the self must distinguish simultaneous phenomenal unity, attribution to one subject, and persistence through change.

Tye argues that experiential unity should not be modeled simply as a collection of independently existing experiences awaiting a further binder. Other approaches appeal to memory connections, embodied activity, causal organization, or a basic subject. Each needs to explain what establishes the boundary of one stream or person.

Split-brain cases complicate the connection between cognitive integration and subject number. Fission thought experiments ask what would happen if one person's psychology continued in two successors. Similarity and continuity alone do not straightforwardly determine which successor is numerically identical with the original. These difficulties affect ego and bundle theories differently, but do not deliver an uncontested result for either.

### Embodiment and Psychological Continuity

Bodily and organism-based approaches locate the continuing individual in a living body. Psychological approaches emphasize memory, character, intentions, and their appropriate causal continuity. The two can disagree about transplantation, amnesia, radical psychological change, and duplication.

An organization-based approach instead asks which processes coordinate perception, correction, planning, and action over time. This supplies possible criteria of cognitive integration, but a particular intervention test or control boundary remains a theoretical proposal, not the accepted definition of the self. Integrated control, biological survival, and phenomenal subjectivity can come apart in both explanations and thought experiments.

### The Experiential Self

Dan Zahavi defends a minimal experiential self grounded in the first-person character of experience. An experience is given to its subject without requiring that subject to reflect on it or make an autobiographical judgment. This **pre-reflective self-awareness** does not posit a little observer inspecting another internal state.

The view offers a reply to a simple bundle argument: the absence of an additional ego-object leaves untouched the first-person manner in which experiences occur. Critics ask whether that manner justifies speaking of a self or merely redescribes the experience. Defenders argue that subjectivity supplies the relevant minimal sense of self, while leaving richer forms of personality and narrative open.

### Narrative and Social Identity

Daniel Dennett's "center of narrative gravity" treats the self as an abstraction organized through stories, interpretations, and social interactions. Like a center of gravity, it can support explanations without being one additional material object hidden among the processes it summarizes.

This conception focuses on the explanatory organization of a person, not a literal narrator inside the brain.

Galen Strawson challenges both the empirical thesis that everyone experiences life narratively and the ethical thesis that a narrative outlook is necessary for living well. His "Episodic" orientation lacks the strong sense of a self stretching through a long personal history while retaining ordinary knowledge of bodily and biographical continuity.

Narrative theorists can respond that implicit organization or social interpretation need not require conscious storytelling. But the broader the category becomes, the more carefully it must distinguish narrative from memory, agency, or any temporal continuity. A minimal experiential subject and a narrative identity may therefore require different accounts rather than competing for a single definition.

**Open Questions**

Questions remain about the relationship between bodily continuity and psychological change, whether experience always involves minimal self-awareness, and how cognitive integration relates to a unified subject. Artificial and distributed systems add boundary problems without making organizational persistence a settled test of personhood. Practical interests and welfare raise further ethical questions; neither automatically identifies the bearer of consciousness.

**Further Reading**

- Tye, Michael. 2003. *Consciousness and Persons: Unity and Identity*. MIT Press.
- Dennett, Daniel C. 1992. "The Self as a Center of Narrative Gravity." In *Self and Consciousness: Multiple Perspectives*, edited by F. Kessel, P. Cole, and D. Johnson. Erlbaum.
- Zahavi, Dan. 2011. "The Experiential Self: Objections and Clarifications." In *Self, No Self?*. Oxford University Press.
- Strawson, Galen. 2008. "Against Narrativity." In *Real Materialism and Other Essays*. Clarendon Press.

**Related Writing on This Site**

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind, matter, and meaning.

**My View**

The self is real, relational, and assembled. I side with the bundle against the ego as an inner thing, while refusing the conclusion that the self is therefore unreal. A persisting integrated control organization can bear content and action without a hidden proprietor; language and public practice can assemble the further autobiographical *I*. This account leaves no role for a non-physical proprietor inside the self; the broader theological question of whether nature has an author remains separate. Treating the self as an immediate inner object reifies it; treating it as a fiction overcorrects. The right register is process, organization, and achievement.

### The Turing Test

The Turing Test is a family of conversational tests inspired by Alan Turing's 1950 "Computing Machinery and Intelligence." Rather than first defining "thinking," Turing proposed an imitation game conducted through written exchanges. The familiar modern version asks whether an interrogator can distinguish a machine from a human conversational partner.

The test provides a way to investigate machine performance without relying on appearance or material composition. Whether success supplies a definition of intelligence, a sufficient condition for thinking, or useful but defeasible evidence remains disputed.

### The Original Imitation Game

Turing begins with a game involving a man, a woman, and an interrogator who tries to identify them through written responses. One participant attempts to mislead the interrogator; the other helps. He then asks what happens if a machine takes the misleading participant's place.

Later formulations center on distinguishing a machine from a human. The relationship between the original sex-identification game and the familiar human-machine test has generated interpretive disagreement. Neither formulation should be silently substituted for the other when attributing the details of the 1950 paper.

The written channel screens off voice, appearance, and physical prowess. Questions can range widely across arithmetic, everyday knowledge, literature, and reasoning. Turing predicts that by the end of the century an average interrogator would have no more than a 70 percent chance of correct identification after five minutes. This was a prediction under particular conditions, not a timeless pass threshold for every later experiment.

### Consciousness and the Problem of Other Minds

Turing discusses an "argument from consciousness" that demands genuinely felt thoughts and emotions rather than mere output. He objects that insisting on direct access to another being's experience would also undermine ordinary judgments about other humans. A conversation can give evidence for mentality even though it does not let the interrogator become the other subject.

He acknowledges mystery about consciousness while arguing that it need not prevent investigation of thinking machines. This is more nuanced than either presenting his test as proof of phenomenal consciousness or saying he simply defines thinking as a permanently deflated substitute for all inner life. His argument concerns what kinds of evidence and standards can reasonably govern attribution.

### Mathematical, Learning, and Originality Objections

Turing considers theological restrictions, mathematical limits on formal machines, and alleged disabilities such as inability to learn or originate anything. Gödel-style limits show constraints on particular formal systems; Turing questions why human beings should be assumed exempt from relevant limits.

"Lady Lovelace's objection" says a machine can do only what people know how to order it to perform. Turing emphasizes that machines can surprise their designers and develops the idea of

learning machines. The broader question is whether programmed origins preclude novelty, not simply whether a programmer supplied an initial mechanism.

**Searle's Chinese Room**

John Searle imagines a person producing appropriate Chinese responses through formal rules without understanding Chinese. He argues that implementing a program alone is insufficient for understanding. Applied to a conversation test, the scenario challenges the move from successful responses to semantics.

The systems reply distinguishes the operator from the whole room. Its defenders argue that the larger organized system could understand even if one component does not. Searle attempts to answer by imagining the operator internalizing the entire arrangement. Whether that reply establishes the absence of system-level understanding remains contested; the Chinese Room is an argument against sufficiency, not an agreed counterexample.

**Block's Lookup-Table Challenge**

Ned Block considers a vast table of appropriate responses for every permitted conversational history in a bounded test. Its behavior could match a person even though its information-processing organization differs radically from an ordinary thinker. Block argues that intelligence depends partly on how behavior is produced.

Such a device would still execute a computational lookup. The objection does not require pretending it computes nothing. It asks whether the computation supports the intelligence attributed on the basis of the transcript. Test defenders question the relevance of astronomically impractical tables and emphasize the evidential force of flexible success across unpredictable exchanges.

**Language Models and Test Design**

Language models make the distinction between conversational fluency and its underlying mechanisms practically important. However, there is no single standardized test whose result licenses claims about every system. Prompting, time limits, comparison humans, judges, sampling, tools, and incentives all affect outcomes. Humanlike writing alone is not a complete report of performance in Turing's proposed game.

Emily Bender and Alexander Koller argue that linguistic form alone does not establish meaning and communicative intent. Other accounts emphasize information inherited from human practice, learned internal representations, or the added capacities of multimodal and tool-using systems. Their dispute concerns the significance of training and organization, not whether fluent output occurs.

Success can provide evidence of conversational competence without independently establishing consciousness, general understanding, or personhood. Failure can also reflect unfamiliarity with human conversational conventions rather than an absence of every kind of intelligence. The open question is how behavioral evidence should combine with evidence about architecture, learning, and interaction.

**Further Reading**

- Turing, Alan M. 1950. "Computing Machinery and Intelligence." *Mind* 59: 433–460.
- Block, Ned. 1981. "Psychologism and Behaviorism." *Philosophical Review* 90: 5–43.
- Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3: 417–424.
- Oppy, Graham, and David Dowe. 2021 revision. "The Turing Test." *Stanford Encyclopedia of Philosophy*.
- Bender, Emily M., and Alexander Koller. 2020. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198.

**Related Writing on This Site**

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind, matter, and meaning.

**My View**

The Turing Test sets exactly the bar my view rejects — and it is worth being precise about where the fault lies, because it does not lie with Turing.

In 1950, with behaviorism ascendant and no science of the inner life in prospect, an operational test was a reasonable, clarifying move, and Turing was scrupulous: he *bracketed* consciousness rather than pretending his test disposed of it. Give him that. The error enters with his successors, who forget the bracketing and brandish a passed test as proof of understanding or experience. The test's own author declined that inference. So do I.

My objection to the criterion is the objection that sank behaviorism, of which the test is the cleanest single instrument. Behavior alone underdetermines the organization producing it. Conversational performance can supply defeasible evidence of content-guided competence, but a terminal transcript cannot by itself identify the bearer, reveal content-fixing history, show that several capacities compose understanding, or establish phenomenal organization. Those questions require evidence about the system behind the exchange. See *The Symbol Grounding Problem* and *Semantic Externalism*.

The LLM moment therefore vindicates the objection without making performance worthless. Winning the game and having a mind remain different achievements; success at one can still count as evidence relevant to the other. The anthropomorphic reflex makes fluent talk feel more conclusive than it is, while provenance and architecture determine how much evidential weight survives. Turing named the modern question well. His readers erred when they turned one behavioral measure into a verdict about every dimension of mind.

**Transparency**

Transparency concerns the apparent world-directedness of introspection on perceptual experience. When someone attends to seeing a red apple, attention ordinarily falls on the apparent apple, its color, and its shape, rather than a second private surface carrying those qualities. The experience can seem transparent to what it presents.

Philosophers disagree about the strength and scope of this observation. A modest thesis says introspection depends on attending to represented objects and properties. Stronger versions deny direct awareness of intrinsic properties of experience. Neither formulation automatically proves that phenomenal character is identical with representational content.

**Historical Motivation**

G. E. Moore's discussion of the apparently diaphanous character of sensation supplies an early point of reference. Gilbert Harman's "The Intrinsic Quality of Experience" makes the observation central to a challenge against treating experienced colors as intrinsic qualities of mental states. Michael Tye develops it into an argument for representationalism.

This history does not make transparency exclusively representationalist. As Tye notes, a sense-datum theorist can distinguish the act of sensing from the thing sensed and assign the noticed qualities to the latter. Moore's diaphanous sensation therefore should not be read as already endorsing an external-object theory of perceptual content.

Consider blue seen on an ocean's surface. The blue appears to belong to the surface, not to the neural process enabling sight. An experience representing blue need not itself be blue, just as a representation of a large object need not have that object's size. This vehicle/content distinction explains why the location and physical properties of the experience's vehicle cannot simply be read from what appears.

**Tye's Argument from Transparency**

In "Representationalism and the Transparency of Experience," Tye develops ten steps. The first seven concern vision; the last three extend the approach. Their central moves can be summarized as follows:

1. Visual qualities appear as qualities of presented surfaces, rather than introspectively accessible qualities of the experience.
2. Treating them instead as mislocated private qualities threatens to attribute extensive error to ordinary visual presentation.
3. Introspective attention to the experience concerns those presented qualities, not an independently encountered inner display.
4. Hallucination can have a similar character without an actual external object bearing the apparent qualities.
5. Changes in the qualities of which the subject is directly aware necessarily change visual phenomenal character.
6. Representational content of an appropriate kind offers, on Tye's account, the best explanation of these facts.
7. This rejects qualia defined as directly accessible intrinsic qualities of experiences, while preserving qualities contributing to what experience is like.
8. The account extends to hearing, smell, and taste: the directly presented qualities appear to belong to sounds, odors, and tastes, not to the experiences.
9. Bodily sensations receive a parallel treatment. A pain presents a disagreeable condition as located in a bodily region; a phantom pain shows that the apparent region need not exist.

10. Felt emotions and moods can present bodily changes or a change in oneself and the surrounding world. Tye argues that these appearances likewise do not reveal qualities of the experience's vehicle.

Step six is explanatory, not a deductively established identity. Dependence on presented qualities alone does not show that no additional property contributes to phenomenal character. Step four also assumes an appropriate common phenomenal type across perception and hallucination, which disjunctivists contest.

**Blurry Vision**

An object can be seen blurrily without being seen as blurry. Unfocusing one's eyes on a sharp-edged printed letter changes the experience without making the letter appear to possess intrinsically fuzzy edges. Paul Boghossian and David Velleman, and Ned Block in related discussions, use this contrast to challenge an exclusively content-based account.

Tye replies that blurry vision can involve less determinate representation of boundaries, whereas seeing an object as blurry represents its edges as objectively indistinct. He distinguishes seeing something blurrily, seeing it as blurry, and seeing a blurry object clearly. In his usage, the second involves a conceptual classification; the first and third need not. The proposed content differences would explain phenomenal differences without an extra inner quality.

A watercolor example makes the contrast concrete. A painting made on wet paper can have genuinely blurred edges seen clearly through glasses. A painting with sharp edges, viewed without glasses, can instead be seen blurrily. Tye reports Frank Jackson's suggestion that the first experience precisely represents an indefinite boundary, whereas the second fails to specify an otherwise definite boundary. The theory distinguishes representing indeterminacy in the object from representing the object indeterminately.

Critics ask whether indeterminacy captures the character of blur or merely its informational consequences. More general intentionalist approaches distinguish representational content from the manner or mode of presentation. The question is whether that additional structure remains compatible with the intended transparency thesis.

**Inversion and Hallucination**

Inverted-spectrum cases imagine matching representations with different phenomenal character. If possible, such cases would threaten strong representationalism. They need not refute the modest observation that attention returns to the presented scene. Representationalists challenge the possibility or the description of the matching content; critics argue that transparency alone leaves the inversion open.

Naive realist and disjunctivist theories can also accept aspects of transparency. A veridical experience may relate a subject directly to an external object rather than represent a common experiential kind also available in hallucination. Consequently, transparency does not independently choose between representationalism and every form of direct realism.

### Reflective Access and Other Modalities

Tye's 2002 account distinguishes direct awareness of presented qualities from awareness that one is undergoing an experience with a certain character. Introspection is not a scan of the neural vehicle. He models it as a reliable process leading from awareness of the presented qualities to a judgment about one's experience, without conscious inference. Critics can accept the world-directed starting point while disputing the reliability, exhaustiveness, or metaphysical conclusions of that account.

Tim Crane questions whether introspection can settle the correct theory of phenomenal character. Knowing what an experience is like differs from knowing its complete metaphysical nature. Attention can reveal contrasts in clarity, organization, or manner of presentation even when it does not reveal a private colored surface.

Pain, emotion, mood, and imagery further complicate generalization. Pain can present a bodily disturbance rather than an ordinary visible object. A mood may have diffuse world-directed character without selecting one clear target. Whether these experiences are transparent in the same sense as vision remains disputed.

The evidential value of transparency therefore depends on the thesis defended. It can motivate resistance to an inner-display model while leaving questions about consciousness, subjectivity, phenomenal properties, and explanatory identity unanswered. A strong theory must defend those further claims through arguments beyond the original introspective observation.

### Further Reading


- Harman, Gilbert. 1990. "The Intrinsic Quality of Experience." *Philosophical Perspectives* 4: 31–52.
- Crane, Tim. 2000. "Introspection, Intentionality, and the Transparency of Experience." *Philosophical Topics* 28: 49–67.
- Tye, Michael. 2002. "Representationalism and the Transparency of Experience." *Noûs* 36: 137–151.


### Related Writing on This Site

- Mind, Matter, and Meaning — Gordon Swobe's book-length argument about mind, matter, and meaning.

### My View

My view gives transparency a leading but limited evidential role. The datum counts strongly against an inner-object theory: attention does not disclose a private surface whose colors and shapes stand in for the world. It also removes direct acquaintance with mental paint as evidence. It does not deductively establish that no hidden residue exists, and it does not by itself tell us what experience consists in.

The identity claim enters as the best explanation of the wider case. Within a state already counted as conscious and borne by an identified subject, phenomenal character consists in the complete world-presenting intentional profile: independently fixed content under a perceptual or bodily-affective mode, with determinacy, field structure, and attentional organization. That conclusion

depends on the profile being specified independently and on the failure of mental paint to explain any stable phenomenal difference the profile leaves out.

Transparency therefore opens the case rather than closing every debt. Content fixing, state consciousness, and subject individuation remain separate questions. Levine's epistemic gap remains real, and the deflation of the hard problem stays hostage to the identity claim. The phenomenology remains; the inner exhibit loses its advertised evidence.

### Vehicle and Content

The **vehicle** of a representation is what carries it: ink on a page, pixels on a screen, a spoken sound, or a pattern of neural activity. Its **content** is what it represents. A photograph on a desk can depict the Eiffel Tower in Paris; neither the tower's height nor its location is a property of the paper. The distinction separates questions about a representation's physical realization from questions about what it says or depicts.

### Vehicles, Contents, and Objects

Three terms need to be distinguished. The vehicle carries the representation. The content specifies how things are represented. The object, where there is one, is what the representation concerns. A photograph may depict the Eiffel Tower from below. The tower is its object; the way the tower is depicted belongs to its content. A drawing of an imaginary tower has a vehicle and content even if no such tower exists.

Philosophers disagree about the nature of content. Propositional theories treat the content of a belief as something that can be true or false. Other theories describe perceptual content through accuracy conditions, spatial structure, or nonconceptual presentations of properties. Some views make particular objects constituents of content; others distinguish an object's contribution to reference from a more general descriptive content. The vehicle/content distinction does not settle these disputes.

The same content can often be carried by different vehicles: a spoken sentence, its written version, and a translated sentence may express the same claim. Conversely, an identical mark can carry different contents in different languages or contexts. This does not make the vehicle irrelevant. Its structure, history, and use can help explain why it carries one content rather than another.

### Physical Properties and Represented Properties

A map represents a winding road without its paper winding through the countryside. A neural state can represent a red apple without itself being red or apple-shaped. Moving a map to another room need not change the place it represents. These examples show why an inference from the location or properties of a vehicle to the location or properties of its object requires further premises.

Fred Dretske develops the distinction in *Naturalizing the Mind* (1995). For his representational theory, the relevant sensory qualities are properties represented as belonging to objects, rather than qualities that the neural vehicle literally acquires. This helps explain how physical processes can concern things outside the nervous system.

It does not follow that every indirect theory of perception confuses a vehicle with an object. A sense-datum theorist may deliberately posit immediately perceived objects on independent grounds, such as arguments from illusion or hallucination. Nor does every defender of qualia hold that neural tissue literally turns red. Their claim may concern properties of experience that neither resemble the neural vehicle's ordinary physical properties nor reduce to its representational content. Those positions require substantive evaluation, not just a reminder of terminology.

### Perception, Transparency, and Hallucination

Representational theories of perception distinguish the processes enabling vision from what vision presents. Looking at an apple ordinarily directs attention toward its color, shape, and position, not toward the neurons enabling the experience. The **transparency** thesis generalizes this observation: introspection of perceptual experience seems to disclose represented features rather than a further inner object.

Different transparency claims have different strengths. The inability to attend to one's neural vehicle does not establish that every phenomenal feature reduces to content. Tim Crane emphasizes the intentional **mode** under which content is presented, while Ned Block argues that qualitative character can exceed representational content. Seeing and imagining the same scene, or seeing it sharply and blurrily, provide tests for accounts of how content, mode, and phenomenal character relate.

On a representational account, hallucination can involve content whose accuracy conditions are not met. An experience as of a red apple need not involve an actual apple or an inner red object. Relational and disjunctivist theories question whether veridical perception and an indistinguishable hallucination share the same fundamental nature. The vehicle/content distinction remains useful in articulating the alternatives, but does not select between them.

### Accuracy and Perceptual Contact

Accurate content is not always enough for perception of an object. Imagine an actual hand hidden behind a screen. Brain stimulation produces an experience as of a hand at exactly that location, independently of the hidden hand. The experience matches the scene, yet the match alone does not establish that the subject sees that hand.

Causal theories require an appropriate connection between the object and the perceptual process, not merely correspondence between content and reality. They must also explain which causal chains count as appropriate: an object might cause an experience through an unusual route that still fails to constitute seeing it. Relational theories instead make the perceptual relation central to the nature of the experience. The example separates accurate representation from successful perceptual contact.

### Cognitive Science and Artificial Systems

In cognitive science, identifying a neural or computational vehicle differs from determining its content. Researchers may ask which pattern encodes a location, how downstream systems use it, and whether intervention changes behavior in ways predicted by that interpretation. Causal,

informational, teleosemantic, and interpretation-based theories offer different accounts of what makes these assignments correct.

A model's token vectors or weight matrices are candidate vehicles; their contents cannot simply be read off their numerical shape. Training data, learning objectives, environmental relations, and subsequent use may all be relevant. Establishing content at one level does not by itself settle whether a whole system understands language or has conscious experience. Equally, the distinction supplies no general argument that artificial vehicles cannot carry genuine content.

**Further Reading**


- Dretske, Fred. *Naturalizing the Mind*. MIT Press, 1995.
- Crane, Tim. *Elements of Mind: An Introduction to the Philosophy of Mind*. Oxford University Press, 2001.
- Block, Ned. "Mental Paint." In *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, edited by Martin Hahn and Bjørn Ramberg, 165–200. MIT Press, 2003.


**Related Writing on This Site**

- *Mind, Matter, and Meaning* — Gordon Swobe's book-length argument about mind, consciousness, and meaning; an author position, not an endorsement by this reference page or its subjects.

**My View**

My view uses the vehicle/content distinction throughout. Three deployments matter most.

**Transparency.** Intrinsic vehicle properties do not appear as a second colored surface. Within a conscious state borne by an identified subject, phenomenal character consists in the complete world-presenting profile — content under a mode, with its determinacy, field structure, and attentional organization. Transparency supports that claim without settling state consciousness or subjecthood by itself.

**Representation without a representative.** Representation does not require an inner representative — a mental object that stands for the world. It requires a physical vehicle whose history and producer-consumer organization fix accuracy conditions. Content is no further item interposed between subject and world.

**AI and computation.** Token patterns and weight matrices are vehicles. Their geometry alone does not fix content, but their training history and producer-consumer use can help explain what they represent. Mallory gives a serious account of word2vec content about actual word-types and linguistic distributions; originality at that grain remains unsettled. The inherited vocabulary and task do not by themselves establish derivation. Multimodal or interactive histories can add further content-fixing relations without serving as a prerequisite for all original artificial content. Whether mechanism-level content guides domain-specific competence, supports partial or more integrated world-involving understanding, enters phenomenal consciousness, or belongs to a persisting subject must be assessed separately for a named candidate system.

### Zombies and Conceivability

A **philosophical zombie** is a hypothetical duplicate of a conscious person that lacks phenomenal consciousness: there is nothing it is like to be that duplicate. A zombie would say that a tooth hurts and respond as a conscious person does, while experiencing no pain. The thought experiment asks whether all the physical facts could hold without the phenomenal facts, and what our ability to conceive such a situation tells us about reality.

### Physical and Functional Duplicates

A behavioral duplicate merely acts like a person. A functional duplicate shares a specified organization of inputs, internal transitions, and outputs, but may have a different physical realization. A physical duplicate shares the relevant physical facts, including whatever environmental and historical facts a complete physical specification requires. These are different hypotheses.

The argument against physicalism needs the physical-duplicate version. A robot that behaves like a person but uses different mechanisms is not, just for that reason, a counterexample to physicalism. Absent experience in a functionally identical system would instead challenge the sufficiency of that functional organization for consciousness. The relevant functional specification must not already define its states in terms of felt experience, or absence of experience would be excluded by definition.

David Chalmers develops the physical-duplicate argument in *The Conscious Mind* (1996) and later work on modal epistemology, the study of our knowledge of possibility and necessity. His claim concerns **metaphysical possibility**, not whether zombies could arise under the actual laws of nature. Naturalistic dualism can hold that psychophysical laws connect physical processes and consciousness in our world while denying that this connection holds in every possible world.

### The Argument

A simplified version has three stages:

1. A complete physical description together with the absence of consciousness is ideally conceivable.
2. The relevant kind of ideal conceivability establishes a corresponding possibility.
3. If a complete physical duplicate can lack consciousness, physical facts do not metaphysically necessitate phenomenal facts, contrary to physicalism.

Each stage needs qualification. Ordinary imaginability is too weak for the second premise, and more technical versions distinguish different ways of evaluating a possible scenario. Chalmers's two-dimensional argument also considers whether the intrinsic nature underlying physical structure could yield a Russellian alternative to standard materialism. The simple formulation identifies the issue; it is not the full argument.

### Varieties of Conceivability

Chalmers distinguishes three dimensions:

- **Prima facie and ideal:** a scenario may seem coherent initially yet fail under better reasoning. Ideal conceivability survives ideal rational reflection.

- **Negative and positive:** inability to rule a statement out a priori differs from constructing a coherent scenario that verifies it.
- **Primary and secondary:** a scenario may be considered as a hypothesis about how actuality is, or as a counterfactual alternative evaluated with actual reference fixed.

The last distinction addresses necessary truths discovered empirically. Before learning its chemical composition, one can consider a scenario in which the liquid in the oceans is XYZ rather than $H_2O$. Considered as actual, that scenario could verify that water is XYZ. Considered counterfactually after actual water has been identified as $H_2O$, XYZ would be another liquid, not water. Thus primary conceivability does not automatically establish secondary possibility for the same sentence.

Chalmers argues that core phenomenal concepts differ from "water" in the relevant respect. There is no analogous separation between pain and a merely pain-like feeling: if something genuinely feels painful, it is a pain experience. He uses this to defend the further steps from primary zombie conceivability toward the metaphysical conclusion. Both the adequacy of this treatment of phenomenal concepts and the broader modal principles remain contested.

### Replies and Counter-Replies

#### Rejecting the conceivability premise

Some physicalists hold that a sufficiently complete understanding of physical or functional facts would make zombie descriptions incoherent. On this approach, we mistake a gap in our present understanding for the coherence of a fully specified scenario. Functionalists and representationalists may argue that holding the relevant organization fixed while removing experience fails to preserve what their theories identify as experience.

Chalmers replies that asserting the identity under dispute does not show an a priori contradiction in the zombie description. A rebuttal must explain why ideal reflection rules the scenario out, rather than assuming physicalism as a premise. Disputes about whether imagining "no experience" actually specifies a coherent world also bear on positive conceivability.

#### Accepting an epistemic gap without a modal gap

Type-B physicalists accept an epistemic gap between physical and phenomenal descriptions while treating their connection as metaphysically necessary. Phenomenal-concept strategies explain the gap through distinctive ways of thinking about experience: recognition, acquaintance, or first-person access may yield concepts cognitively independent of descriptive physical concepts even if they concern the same property.

Chalmers's objection is that explaining cognitive independence does not yet explain why ideal primary conceivability fails to indicate possibility. He argues that the reply requires **strong necessities**, necessary connections unavailable a priori that lack the ordinary two-dimensional explanation of necessary a posteriori truths. Physicalists dispute whether those necessities are objectionable, whether phenomenal concepts have the structure Chalmers assigns them, and whether his modal framework is compulsory.

### Questioning the methodology

Other critics question whether intuitions about such remote scenarios provide reliable evidence about consciousness. Defenders answer that modal reasoning also operates elsewhere in philosophy and should be assessed by its best formulation, not identified with casual imagination. The dispute concerns the justification of the inference, not whether an unusual scenario feels intuitive to every reader.

### What the Thought Experiment Establishes

Zombie conceivability, zombie possibility, and evidence for an actual unconscious system are separate matters. Conceivability alone is not empirical evidence that any person or machine lacks experience. If the required metaphysical possibility is established, it poses a direct problem for physical necessitation of consciousness. Whether it is established depends on the conceivability premise, the modal bridge, and the interpretation of physical and phenomenal concepts.

### Further Reading


- Chalmers, David J. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, 1996.
- Chalmers, David J. "Does Conceivability Entail Possibility?" In *Conceivability and Possibility*, edited by Tamar Szabó Gendler and John Hawthorne, 145–200. Oxford University Press, 2002.
- Chalmers, David J. "Consciousness and Its Place in Nature." In *The Blackwell Guide to Philosophy of Mind*, edited by Stephen P. Stich and Ted A. Warfield. Blackwell, 2003.
- Block, Ned. "Troubles with Functionalism." *Minnesota Studies in the Philosophy of Science* 9 (1978): 261–325.


### Related Writing on This Site

- *Mind, Matter, and Meaning* — Gordon Swobe's book-length argument about mind, consciousness, and meaning; an author position, not an endorsement by this reference page or its subjects.

### My View

The zombie argument is the most prominent contemporary anti-physicalist argument, and I treat it as a structural opponent. Of the three replies above, my weight falls on Line 3 — the type-B response — with Line 2 in support. Line 1 I decline, and the reason matters: telling a reader that the thing they can plainly do, they cannot really do, spends credibility I would rather keep. The inner-theater diagnosis still does heavy work on this page and throughout the book, but as an account of *why* the conceiving comes so easily, never as a denial that it comes.

The position: zombies are conceivable, ideally and permanently. I take the conceivability premise off the table as a concession, not a battle. What I refuse is the bridge from ideal conceivability to metaphysical possibility.

The acquaintance-recognition account explains why the conceiving persists if the identity claim holds. For knowers with our architecture, undergoing an experience grounds a recognitional capacity that no public description by itself confers. The acquaintance-grounded and descriptive routes can therefore remain permanently separable in thought while reaching one physical fact. That makes ideal zombie conceivability compatible with Type-B physicalism. It does not establish

the identity, remove the conceiving's possible modal force, or show that every possible physical knower must share our conceptual architecture.

Chalmers's reply marks the remaining dispute. Core phenomenal concepts, he argues, lack the contingent reference-fixing story that turns water without $H_2O$ into a merely apparent possibility. Their ideal primary conceivability therefore reaches possibility on his view. A Type-B physicalist must accept a strong necessity with no *a priori* route between its terms. I accept that revisionary cost and deny the further principle that every such conceptual separation maps onto a possible world. Acquaintance-recognition explains why the gap persists; the identity claim and the strong-necessity dispute carry the modal burden. Neither side gets that verdict free.

# Works Cited


- *Merriam-Webster's Collegiate Dictionary*, 11th ed. (Springfield, MA: Merriam-Webster, 2003). Ch05 n.2, Ch06 n.1
- Scott Aaronson. "Why I Am Not An Integrated Information Theorist (or, The Unconscious Expander)." *Shtetl-Optimized* (blog), May 21, 2014. App. B n.62
- Robert M. Adams. *Leibniz: Determinist, Theist, Idealist* (New York: Oxford University Press, 1994). Ch12 n.8
- Ekin Akyürek et al. "What Learning Algorithm Is In-Context Learning? Investigations with Linear Models." arXiv:2211.15661 (2022). Ch23 n.10b
- Larissa Albantakis et al. "Integrated Information Theory (IIT) 4.0: Formulating the Properties of Phenomenal Existence in Physical Terms." *PLOS Computational Biology* 19, no. 10 (2023): e1011465. DOI 10.1371/journal.pcbi.1011465. App. B n.60
- Michael V. Antony. "Are Our Concepts CONSCIOUS STATE and CONSCIOUS CREATURE Vague?" *Erkenntnis* 68, no. 2 (2008): 239–263. Ch10 n.7
- Michael V. Antony. "Papineau on the Vagueness of Phenomenal Concepts." *Dialectica* 60, no. 4 (2006): 475–483. DOI 10.1111/j.1746-8361.2006.01083.x. App. B n.40
- Michael V. Antony. "Vagueness and the Metaphysics of Consciousness." *Philosophical Studies* 128, no. 3 (2006): 515–538. DOI 10.1007/s11098-004-7488-8. Ch10 n.7, App. B n.40
- Thomas Aquinas. *Summa theologiae*. I, q. 85, a. 2. Trans. Fathers of the English Dominican Province, second revised ed. (1920). Online edition, New Advent (2017). https://www.newadvent.org/summa/1085.htm. Ch06 n.3
- D. M. Armstrong. *A Materialist Theory of the Mind* (London: Routledge & Kegan Paul, 1968). Ch18 n.7
- J. L. Austin. *How to Do Things with Words*, ed. J. O. Urmson (Oxford: Clarendon Press, 1962). Ch16 n.11
- Edward W. Averill and Joseph Gottlieb. "Two Theories of Transparency." *Erkenntnis* 86, no. 3 (2021): 553–573. Transparency n.10
- Anita Avramides. *Other Minds* (Routledge, 2001). Ch19 n.1
- Murat Aydede and Güven Güzeldere. "Some Foundational Problems in the Scientific Study of Pain." *Philosophy of Science* 69, no. S3 (2002): S265–S283. DOI 10.1086/341851. Ch07 n.7
- Murat Aydede. "Is Feeling Pain the Perception of Something?" *Journal of Philosophy* 106, no. 10 (2009): 531–567. DOI 10.5840/jphil20091061033. Ch07 n.7
- A. J. Ayer. *The Foundations of Empirical Knowledge* (London: Macmillan, 1940). Ch03 n.2, Argument from Illusion nn.1, 10, Ch03 epigraph
- Michael Ayers. *Locke*. 2 vols.: *Epistemology* and *Ontology* (London: Routledge, 1991). Ch02 n.4

- David Bain. "Intentionalism and Pain." *The Philosophical Quarterly* 53, no. 213 (2003): 502–523. Ch07 n.4
- David Bain. "What Makes Pains Unpleasant?" *Philosophical Studies* 166 (2013): S69–S89. DOI 10.1007/s11098-012-0049-7. Ch07 n.4
- Lynne Rudder Baker. *Persons and Bodies: A Constitution View* (Cambridge: Cambridge University Press, 2000). Ch06 n.21, Ch06 n.21b
- Katalin Balog. "In Defense of the Phenomenal Concept Strategy." *Philosophy and Phenomenological Research* 84, no. 1 (2012): 1–23. DOI 10.1111/j.1933-1592.2011.00541.x. App. B n.17
- Justin L. Barrett. "Exploring the Natural Foundations of Religion." *Trends in Cognitive Sciences* 4, no. 1 (2000): 29–34. doi:10.1016/S1364-6613(99)01419-9. Ch25 n.7
- Justin L. Barrett. *Why Would Anyone Believe in God?* (Walnut Creek, CA: AltaMira Press, 2004). Ch25 n.7
- Friedrich Beck and John C. Eccles. "Quantum Aspects of Brain Activity and the Role of Consciousness." *Proceedings of the National Academy of Sciences* 89, no. 23 (1992): 11357–11361. DOI 10.1073/pnas.89.23.11357. Ch09 n.7
- Paul Benacerraf. "What Numbers Could Not Be." *The Philosophical Review* 74, no. 1 (1965): 47–73. DOI 10.2307/2183530. Ch12 n.7
- Emily M. Bender and Alexander Koller. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020): 5185–5198. Ch23 n.1, Ch24 n.1, Ch24 n.14, Ch25 n.1, Ch25 n.11, App. B n.4
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021), 610–623. DOI 10.1145/3442188.3445922. Ch21 n.7
- José Luis Bermúdez. *Thinking Without Words* (Oxford: Oxford University Press, 2003). Ch17 n.2
- Marcelo Berthier, Sergio Starkstein, and Ramón Leiguarda. "Asymbolia for Pain: A Sensory-Limbic Disconnection Syndrome." *Annals of Neurology* 24, no. 1 (1988): 41–49. DOI 10.1002/ana.410240109. Ch07 n.6
- Mark H. Bickhard. "Representational Content in Humans and Machines." *Journal of Experimental & Theoretical Artificial Intelligence* 5, no. 4 (1993): 285–333. Ch15 n.9a
- Mark H. Bickhard. "The Interactivist Model." *Synthese* 166, no. 3 (2009): 547–591. DOI 10.1007/s11229-008-9375-x. Ch15 n.9a
- Jonathan Birch. *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI* (Oxford: Oxford University Press, 2024). Coda n.4

- Max Black. "The Identity of Indiscernibles." *Mind* 61, no. 242 (1952): 153–164. DOI 10.1093/mind/LXI.242.153. Ch12 n.3
- Ned Block. "Advertisement for a Semantics for Psychology." *Midwest Studies in Philosophy* 10 (1986): 615–678. doi:10.1111/j.1475-4975.1987.tb00558.x. Ch20 n.12, Ch23 n.4
- Ned Block. "Can Only Meat Machines Be Conscious?" *Trends in Cognitive Sciences* 30, no. 4 (2026): 298–308. DOI 10.1016/j.tics.2025.08.009. Ch24 n.15
- Ned Block. "Consciousness, Accessibility, and the Mesh between Psychology and Neuroscience." *Behavioral and Brain Sciences* 30, nos. 5–6 (2007): 481–499. DOI 10.1017/S0140525X07002786. App. B n.75
- Ned Block. "If Perception Is Probabilistic, Why Does It Not Seem Probabilistic?" *Philosophical Transactions of the Royal Society B* 373, no. 1755 (2018): 20170341. DOI 10.1098/rstb.2017.0341. Ch02 n.7
- Ned Block. "Inverted Earth." *Philosophical Perspectives* 4 (1990): 53–79. Ch02 n.10, Ch06 nn.15, 18a, Ch20 n.10, Inverted Earth nn.1, 7
- Ned Block. "Mental Paint and Mental Latex." In *Philosophical Issues*, vol. 7, ed. Enrique Villanueva (Atascadero, CA: Ridgeview, 1996), 19–49. Ch04 n.5
- Ned Block. "Mental Paint." In *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, ed. Martin Hahn and Bjørn Ramberg (Cambridge, MA: MIT Press, 2003), 165–200. Ch04 nn.5–6, Ch05 n.9, Ch06 nn.5–6, 10, 15, Ch07 n.7, Ch11 n.19, Ch15 n.13, Inverted Earth nn.1, 7–8, 10, 14, Transparency n.3, App. B n.38
- Ned Block. "On a Confusion about a Function of Consciousness." *Behavioral and Brain Sciences* 18, no. 2 (1995): 227–247. Ch22 n.14, Ch24 n.15, Inverted Earth n.5
- Ned Block. "The Higher Order Approach to Consciousness Is Defunct." *Analysis* 71, no. 3 (2011): 419–431. Ch06 n.7b, App. B n.58
- Ned Block. "The Mind as the Software of the Brain." In *Thinking: An Invitation to Cognitive Science*, 2nd ed., vol. 3, ed. Edward E. Smith and Daniel N. Osherson, 377–425 (Cambridge, MA: MIT Press, 1995). DOI 10.7551/mitpress/3966.003.0016. Ch21 n.1
- Ned Block. "Troubles with Functionalism." In *Perception and Cognition: Issues in the Foundations of Psychology*, ed. C. Wade Savage, *Minnesota Studies in the Philosophy of Science* 9 (Minneapolis: University of Minnesota Press, 1978), 261–325. Ch20 n.1, Ch20 n.13, Ch20 n.14, Ch22 n.13, App. B n.41
- Ned Block. "Psychologism and Behaviorism." *The Philosophical Review* 90, no. 1 (1981): 5–43. DOI: 10.2307/2184371. Ch24 n.1
- Paul A. Boghossian and J. David Velleman. "Colour as a Secondary Quality." *Mind* 98, no. 389 (1989): 81–103. Ch04 n.6, Ch06 n.11, App. B nn.22, 23

- Princess Elisabeth of Bohemia and René Descartes. *The Correspondence between Princess Elisabeth of Bohemia and René Descartes*, ed. and trans. Lisa Shapiro (Chicago: University of Chicago Press, 2007). Ch09 n.6
- Gemma Boleda. "Distributional Semantics and Linguistic Theory." *Annual Review of Linguistics* 6 (2020): 213–234. DOI 10.1146/annurev-linguistics-011619-030303. Ch23 n.5
- Davide Bordini. "Introspection and the Transparency of Experience." In *The Routledge Handbook of Introspection*, ed. Anna Giustina (London: Routledge, 2026), 373–389. doi:10.4324/9781003393269-31. Transparency n.4, Transparency n.10
- Marcelino Botin. "Russellian Physicalists Get Our Phenomenal Concepts Wrong." *Philosophical Studies* 180, no. 7 (2023): 1829–1848. DOI 10.1007/s11098-023-01955-1. Ch10 n.11
- Pascal Boyer. *Religion Explained: The Evolutionary Origins of Religious Thought* (New York: Basic Books, 2001). Ch25 n.7
- Robert B. Brandom. *Making It Explicit: Reasoning, Representing, and Discursive Commitment* (Cambridge, MA: Harvard University Press, 1994). Ch16 n.17, Ch23 n.4
- Franz Brentano. *Psychology from an Empirical Standpoint*, ed. Oskar Kraus, English ed. Linda L. McAlister, trans. Antos C. Rancurello, D. B. Terrell, and Linda L. McAlister, with an introduction by Peter Simons (London: Routledge, 1995 [1874]). Ch01 n.1, Ch06 n.3, Ch13 n.4, Ch18 n.7, App. B n.28, Ch01 epigraph
- Bill Brewer. "How to Account for Illusion." In *Disjunctivism: Perception, Action, Knowledge*, ed. Adrian Haddock and Fiona Macpherson (Oxford: Oxford University Press, 2008), 168–180. DOI 10.1093/acprof:oso/9780199231546.003.0007. Argument from Illusion n.6
- Bill Brewer. "Perception and Content." *European Journal of Philosophy* 14, no. 2 (2006): 165–181. DOI 10.1111/j.1468-0378.2006.00220.x. Ch06 n.18, App. B n.74, Argument from Illusion n.6
- Bill Brewer. *Perception and Its Objects* (Oxford: Oxford University Press, 2011). DOI 10.1093/acprof:oso/9780199260256.001.0001. Ch06 n.18
- C. D. Broad. *Scientific Thought* (London: Kegan Paul, Trench, Trübner & Co., 1923). Ch02 n.6
- Christopher Devlin Brown. "Russellian Physicalism, Phenomenal Concepts, and Revelation." *Philosophia* 51, no. 5 (2023): 2715–2722. DOI 10.1007/s11406-023-00697-y. Ch10 n.11
- Tyler Burge. "Individualism and the Mental." *Midwest Studies in Philosophy* 4 (1979): 73–121. DOI 10.1111/j.1475-4975.1979.tb00374.x.
- Tyler Burge. *Origins of Objectivity* (Oxford: Oxford University Press, 2010). Ch13 n.6, Ch14 n.1
- Patrick Butlin and Theodoros Lappas. "Principles for Responsible AI Consciousness Research." arXiv:2501.07290 (2025). Coda n.9
- Patrick Butlin et al. "Identifying Indicators of Consciousness in AI Systems." *Trends in Cognitive Sciences* 30, no. 6 (2026): 488–501. DOI 10.1016/j.tics.2025.10.011. Ch24 n.7, Coda n.4

- Patrick Butlin, Robert Long, et al. *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*. arXiv:2308.08708v3 (2023). doi:10.48550/arXiv.2308.08708. Ch24 n.7, Coda n.4, App. B nn.67, 76
- Patrick Butlin. "Higher-order Representation in AI." *Philosophy and the Mind Sciences* 7, no. 1 (2026). DOI 10.33735/phimisci.2026.12032. App. B n.69
- Alex Byrne and David R. Hilbert. "Color Realism and Color Science." *Behavioral and Brain Sciences* 26, no. 1 (2003): 3–21. DOI 10.1017/S0140525X03000013. Ch06 n.18b, Ch14 n.7, App. B nn.22, 25, Inverted Earth n.9
- Alex Byrne and Heather Logue, eds. *Disjunctivism: Contemporary Readings* (Cambridge, MA: MIT Press, 2009). Ch03 n.6, Argument from Hallucination n.11
- Alex Byrne and Michael Tye. "Qualia ain't in the Head." *Noûs* 40, no. 2 (2006): 241–255. Ch06 n.18a, Ch15 n.13
- Alex Byrne. "Intentionalism Defended." *The Philosophical Review* 110, no. 2 (2001): 199–240. Ch04 n.3, Ch06 n.4, Transparency n.8
- Alex Byrne. "Inverted Qualia." *Stanford Encyclopedia of Philosophy*, substantive revision January 29, 2025. https://plato.stanford.edu/entries/qualia-inverted/. App. B (inversion discussion)
- John Campbell. "A Simple View of Colour." In *Reality, Representation, and Projection*, ed. John Haldane and Crispin Wright (New York: Oxford University Press, 1993), 257–268. DOI 10.1093/oso/9780195078787.003.0011. Ch06 n.18b, App. B n.24
- Marisa Carrasco, Sam Ling, and Sarah Read. "Attention Alters Appearance." *Nature Neuroscience* 7, no. 3 (2004): 308–313. DOI 10.1038/nn1194. Ch06 n.10a
- Peter Carruthers and Bénédicte Veillet. "The Phenomenal Concept Strategy." *Journal of Consciousness Studies* 14, nos. 9–10 (2007): 212–236. App. B n.35
- David J. Chalmers. "Consciousness and Its Place in Nature." *Philosophy of Mind: Classical and Contemporary Readings* (New York: Oxford University Press, 2002). Preface n.4, Ch10 n.1, Ch11 n.12, App. B n.9, App. B n.11
- David J. Chalmers. "Could a Large Language Model Be Conscious?" *Boston Review*, August 9, 2023.
- David J. Chalmers. "Does a Rock Implement Every Finite-State Automaton?" *Synthese* 108, no. 3 (1996): 309–333. doi:10.1007/BF00413692. Ch20 n.11, Ch22 n.5
- David J. Chalmers. "Does Conceivability Entail Possibility?" In *Conceivability and Possibility*, ed. Tamar Szabó Gendler and John Hawthorne (Oxford: Oxford University Press, 2002), 145–200. Ch09 n.5, Ch11 n.12, App. B nn.11, 12
- David J. Chalmers. "Facing Up to the Problem of Consciousness." *Journal of Consciousness Studies* 2, no. 3 (1995): 200–219. Ch02 nn.10–11, Ch06 n.22, Ch09 epigraph and n.3, Ch11 n.3, Coda n.3

- David J. Chalmers. "Imagination, Indexicality, and Intensions." *Philosophy and Phenomenological Research* 68, no. 1 (2004): 182–190. DOI 10.1111/j.1933-1592.2004.tb00334.x. Ch08 n.13
- David J. Chalmers. "Perception and the Fall from Eden." In *Perceptual Experience*, ed. Tamar Szabó Gendler and John Hawthorne (Oxford: Oxford University Press, 2006), 49–125. App. B nn.23–24
- David J. Chalmers. "Phenomenal Concepts and the Explanatory Gap." In *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, ed. Torin Alter and Sven Walter (New York: Oxford University Press, 2007), 167–194. doi:10.1093/acprof:oso/9780195171655.003.0009. Ch08 n.14, Ch11 nn.8, 18, App. B nn.13, 17, 35
- David J. Chalmers. "The Meta-Problem of Consciousness." *Journal of Consciousness Studies* 25, nos. 9–10 (2018): 6–61. Ch11 n.10, App. B nn.15, 55
- David J. Chalmers. "The Two-Dimensional Argument Against Materialism." In *The Character of Consciousness* (New York: Oxford University Press, 2010), 141–206. DOI 10.1093/acprof:oso/9780195311105.003.0006. Ch11 n.12
- David J. Chalmers. *Reality+: Virtual Worlds and the Problems of Philosophy* (New York: W. W. Norton, 2022). Ch22 n.15
- David J. Chalmers. *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press, 1996). Ch06 n.22, Ch09 nn.3, 8, Ch11 nn.3, 9, Ch20 nn.19, 20, Ch22 n.5, Ch24 n.15, App. B nn.9, 32, Inverted Earth n.18
- François Chollet. "On the Measure of Intelligence." arXiv:1911.01547v2 (2019). Intro n.2
- Ron Chrisley. "Why Everything Doesn't Realize Every Computation." *Minds and Machines* 4, no. 4 (1994): 403–420. DOI 10.1007/BF00974167. Ch22 n.5
- Wayne D. Christensen and Clifford A. Hooker. "An Interactivist-Constructivist Approach to Intelligence: Self-Directed Anticipative Learning." *Philosophical Psychology* 13, no. 1 (2000): 5–45. doi:10.1080/09515080050002717. Ch15 n.9a
- Paul M. Churchland. "Reduction, Qualia, and the Direct Introspection of Brain States." *The Journal of Philosophy* 82, no. 1 (1985): 8–28. DOI 10.2307/2026509. Ch08 n.7
- Andy Clark and David Chalmers. "The Extended Mind." *Analysis* 58, no. 1 (1998): 7–19. DOI 10.1093/analys/58.1.7. Ch15 n.9b
- Andy Clark. *Being There: Putting Brain, Body, and World Together Again* (Cambridge, MA: MIT Press, 1997). Ch15 n.11
- Jonathan Cohen, C. L. Hardin, and Brian P. McLaughlin. "True Colours." *Analysis* 66, no. 4 (2006): 335–340. DOI 10.1093/analys/66.4.335. App. B n.22
- Jonathan Cohen. *The Red and the Real: An Essay on Color Ontology* (Oxford: Oxford University Press, 2009). App. B n.23

- David Cole. “Artificial Intelligence and Personal Identity.” *Synthese* 88, no. 3 (1991): 399–417. doi:10.1007/BF00413555. Ch21 n.3
- David Cole. “The Chinese Room Argument.” *Stanford Encyclopedia of Philosophy*, substantive revision October 23, 2024. Ch21 n.3
- Clara Colombatto and Stephen M. Fleming. “Folk Psychological Attributions of Consciousness to Large Language Models.” *Neuroscience of Consciousness* 2024, no. 1 (2024): niae013. doi:10.1093/nc/niae013. Ch25 n.4
- Earl Conee. “Phenomenal Knowledge.” *Australasian Journal of Philosophy* 72, no. 2 (1994): 136–150. DOI 10.1080/00048409412345971. Ch08 n.11
- Michael Tye. *Consciousness, Color, and Content* (MIT Press, 2000). Ch02 n.9, Ch04 n.6, Ch06 nn.7, 10, 11, 18a, 21, Ch07 n.10, Ch10 n.5, Ch11 n.8, Ch14 n.7, Ch15 n.13, App. B nn.14, 25, 38, 45, Inverted Earth nn.6, 8, 16, Transparency n.8
- Susanna Schellenberg. *The Unity of Perception: Content, Consciousness, Evidence* (Oxford University Press, 2018).
- B. Jack Copeland. “Narrow versus Wide Mechanism: Including a Re-Examination of Turing’s Views on the Mind-Machine Issue.” *The Journal of Philosophy* 97, no. 1 (2000): 5–32. DOI 10.5840/jphil20009716. Ch22 n.10
- B. Jack Copeland. “The Church–Turing Thesis.” *Stanford Encyclopedia of Philosophy*, substantive revision December 18, 2023 (first published January 8, 1997). App. B n.18
- Luca Corti. “Organizational Normativity and Teleology: A Critique.” *Synthese* 202, no. 3 (2023): article 96. DOI 10.1007/s11229-023-04255-7.
- Tim Crane and Craig French. “The Problem of Perception.” *Stanford Encyclopedia of Philosophy* (Fall 2021 edition). Ch05 n.4, Ch05 n.8, Ch06 n.4, Ch06 n.16, Ch23 n.13, Argument from Hallucination n.12
- Tim Crane and Katalin Farkas. “Mental Fact and Mental Fiction.” In *Mental Fictionalism: Philosophical Explorations*, ed. Tamás Demeter, T. Parent, and Adam Toon (Routledge, 2022), 303–319. doi:10.4324/9781003091073-20. App. B n.3
- Tim Crane. “Brentano on Intentionality.” In *The Routledge Handbook of Franz Brentano and the Brentano School*, ed. Uriah Kriegel (London: Routledge, 2017), 41–48. doi:10.4324/9781315776460-5. Ch01 n.1, Ch13 n.4, App. B n.28
- Tim Crane. “Intentionality as the Mark of the Mental.” *Royal Institute of Philosophy Supplement* 43 (1998): 229–251. DOI 10.1017/S1358246100004380. Ch01 n.1, App. B n.26
- Tim Crane. “Intentionality.” *Routledge Encyclopedia of Philosophy*, version 1 (1998). DOI 10.4324/9780415249126-V019-1. Ch01 n.1
- Tim Crane. “Introspection, Intentionality, and the Transparency of Experience.” *Philosophical Topics* 28, no. 2 (2000): 49–67. Ch04 n.6, Ch04 n.7, Ch06 n.10, Ch19 n.3, Transparency n.9

- Tim Crane. “Is There a Perceptual Relation?” In *Perceptual Experience*, ed. Tamar Szabó Gendler and John Hawthorne (Oxford: Oxford University Press, 2006), 126–146.
- Tim Crane. “Mental Substances.” In *Minds and Persons*, ed. Anthony O’Hear, *Royal Institute of Philosophy Supplement* 53 (Cambridge: Cambridge University Press, 2003), 229–250. Ch09 n.7
- Tim Crane. “Subjective Facts.” In *Real Metaphysics*, ed. Hallvard Lillehammer and Gonzalo Rodríguez-Pereyra (London: Routledge, 2003), 68–83. Ch06 n.13
- Tim Crane. “The Intentional Structure of Consciousness.” In *Consciousness: New Philosophical Perspectives*, ed. Quentin Smith and Aleksandar Jokic (Oxford: Oxford University Press, 2003), 33–56. DOI 10.1093/oso/9780199241286.003.0003. Ch02 n.8
- Tim Crane. “The Origins of Qualia.” In *History of the Mind–Body Problem*, ed. Tim Crane and Sarah Patterson (London: Routledge, 2000), 169–194. Ch04 n.2, Inverted Earth n.19
- Tim Crane. “The Significance of Emergence.” In *Physicalism and Its Discontents*, ed. Carl Gillett and Barry Loewer (Cambridge: Cambridge University Press, 2001), 207–224. DOI 10.1017/CBO9780511570797.011. Ch11 n.13b, App. B n.33
- Tim Crane. “Tye on Acquaintance and the Problem of Consciousness.” *Philosophy and Phenomenological Research* 84, no. 1 (2012): 190–198. Ch08 n.11
- Tim Crane. “What Is the Problem of Non-Existence?” *Philosophia* 40, no. 3 (2012): 417–434. doi:10.1007/s11406-011-9354-1. App. B n.28
- Tim Crane. “What Is the Problem of Perception?” *Synthesis Philosophica* 20, no. 2 (2005): 237–264. Argument from Hallucination n.8
- Tim Crane. *The Mechanical Mind: A Philosophical Introduction to Minds, Machines and Mental Representation*, 2nd ed. (London: Routledge, 2003). Ch02 n.9
- Tim Crane. *The Objects of Thought* (Oxford: Oxford University Press, 2013). App. B nn.28, 29, Argument from Hallucination n.5
- Brian Cutter and Michael Tye. “Pains and Reasons: Why It Is Rational to Kill the Messenger.” *The Philosophical Quarterly* 64, no. 256 (2014): 423–433. DOI 10.1093/pq/pqu025. Ch07 n.5
- Brian Cutter and Michael Tye. “Tracking Representationalism and the Painfulness of Pain.” *Philosophical Issues* 21, no. 1 (2011): 90–109. doi:10.1111/j.1533-6077.2011.00199.x. Ch07 n.4
- Donald Davidson. “Knowing One’s Own Mind.” *Proceedings and Addresses of the American Philosophical Association* 60 (1987). App. B n.38
- Donald Davidson. “Truth and Meaning.” *Synthese* 17 (1967): 304–323. DOI 10.1007/BF00485035. Ch16 n.6
- DeepSeek-AI et al. “DeepSeek-V3 Technical Report.” arXiv:2412.19437 (2024). Ch20 n.3
- Stanislas Dehaene. *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts* (New York: Viking, 2014). Ch24 n.15, App. B n.76

- Daniel C. Dennett. “Quining Qualia.” In *Consciousness in Contemporary Science*, ed. A. J. Marcel and E. Bisiach (Oxford: Oxford University Press, 1988), 42–77. App. B n.54
- Daniel C. Dennett. “Real Patterns.” *The Journal of Philosophy* 88, no. 1 (1991): 27–51. Ch14 n.14, Ch17 n.11, Ch25 n.12
- Daniel C. Dennett. *Consciousness Explained* (Boston: Little, Brown, 1991). Ch06 nn.7c, 19a, 20, Ch08 n.4, Ch11 n.13a, Ch17 n.9, Ch24 nn.3, 15, App. B n.23
- Daniel C. Dennett. *Darwin’s Dangerous Idea: Evolution and the Meanings of Life* (New York: Simon & Schuster, 1995). Ch24 n.3
- Daniel C. Dennett. *Elbow Room: The Varieties of Free Will Worth Wanting* (Cambridge, MA: MIT Press, 1984). Ch18 n.5
- Daniel C. Dennett. *From Bacteria to Bach and Back: The Evolution of Minds* (New York: W. W. Norton, 2017).
- Daniel C. Dennett. *The Intentional Stance* (Cambridge, MA: MIT Press, 1987). Ch01 n.2, Ch14 n.14, Ch24 n.3, Ch25 n.12
- Daniel Dennett. “What RoboMary Knows.” In *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, edited by Torin Alter and Sven Walter (Oxford University Press, 2007), 15–31. DOI: 10.1093/acprof:oso/9780195171655.003.0001.
- René Descartes. *Meditations on First Philosophy: With Selections from the Objections and Replies*, trans. and ed. John Cottingham (Cambridge: Cambridge University Press, 1996 [1641]). DOI 10.1017/CBO9780511805028. Ch02 n.2
- René Descartes. *The Philosophical Writings of Descartes*, vol. II, trans. John Cottingham, Robert Stoothoff, and Dugald Murdoch (Cambridge: Cambridge University Press, 1984). Meditation VI consulted through the MIT *Minds and Machines* course excerpt. Ch15 n.11
- David Deutsch. “Quantum Theory, the Church–Turing Principle and the Universal Quantum Computer.” *Proceedings of the Royal Society of London A* 400, no. 1818 (1985): 97–117. DOI 10.1098/rspa.1985.0070. Ch22 n.11
- Esa Díaz-León. “Can Phenomenal Concepts Explain the Epistemic Gap?” *Mind* 119, no. 476 (2010): 933–951. DOI 10.1093/mind/fzq073. App. B n.35
- M. Leann Dodd, Kevin J. Klos, James H. Bower, Yonas E. Geda, Keith A. Josephs, and J. Eric Ahlskog. “Pathological Gambling Caused by Drugs Used to Treat Parkinson Disease.” *Archives of Neurology* 62, no. 9 (2005): 1377–1381. doi:10.1001/archneur.62.9.noc50009. Ch18 n.1
- Adrien Doerig, Aaron Schurger, Kathryn Hess, and Michael H. Herzog. “The Unfolding Argument: Why IIT and Other Causal Structure Theories Cannot Explain Consciousness.” *Consciousness and Cognition* 72 (2019): 49–59. DOI 10.1016/j.concog.2019.04.002. App. B n.60
- Fred Dretske. “Experience as Representation.” *Philosophical Issues* 13, no. 1 (2003): 67–82. DOI 10.1111/1533-6077.00005. Ch13 n.3

- Fred Dretske. "Misrepresentation." In *Belief: Form, Content, and Function*, ed. Radu J. Bogdan (Oxford: Clarendon Press, 1986), 17–36. Ch13 n.1
- Fred Dretske. *Explaining Behavior: Reasons in a World of Causes* (Cambridge, MA: MIT Press, 1988). Ch15 nn.1, 8, Ch18 n.6, Ch22 nn.6, 12, Ch23 n.11, Inverted Earth n.13
- Fred Dretske. *Knowledge and the Flow of Information* (Cambridge, MA: MIT Press, 1981).
- Fred Dretske. *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995). Ch05 n.3, Ch06 n.19, Ch09 n.4, Ch13 n.6, Ch15 nn.8, 13, Ch19 n.3, Ch20 n.18, Ch22 n.12, App. B nn.38, 48, Inverted Earth n.13
- Hubert L. Dreyfus. "Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian." *Philosophical Psychology* 20, no. 2 (2007): 247–268. DOI 10.1080/09515080701239510. Ch20 n.23
- Hubert L. Dreyfus. *What Computers Can't Do: A Critique of Artificial Reason* (New York: Harper & Row, 1972). Ch20 n.22
- Hubert L. Dreyfus. *What Computers Still Can't Do: A Critique of Artificial Reason* (Cambridge, MA: MIT Press, 1992). Ch20 n.22
- Abhimanyu Dubey et al. "The Llama 3 Herd of Models." arXiv:2407.21783v1 (2024). Ch20 n.3
- Leonard Dung and Luke Kersten. "Implementing Artificial Consciousness." *Mind & Language* 40, no. 3 (2025): 285–305. doi:10.1111/mila.12532. Ch24 n.10
- Fallon Durant et al. "Long-Term, Stochastic Editing of Regenerative Anatomy via Targeting Endogenous Bioelectric Gradients." *Biophysical Journal* 112, no. 10 (2017): 2231–2243. DOI 10.1016/j.bpj.2017.04.011. Ch15 n.9c
- Alexandre Duval and Colin Klein. "Pain Asymbolia Is Probably Still Pain." *Philosophy of Science* 93, no. 1 (2026): 221–229. DOI 10.1017/psa.2025.10098. Ch07 n.6
- Andreas Elpidorou. "Phenomenal Concepts." In *Oxford Bibliographies in Philosophy*, ed. Duncan Pritchard (New York: Oxford University Press, 2015). Ch11 n.6
- Evelyn Elsaesser, Chris A. Roe, Callum E. Cooper, and David Lorimer. "The Phenomenology and Impact of Hallucinations Concerning the Deceased." *BJPsych Open* 7, no. 5 (2021): e148. doi:10.1192/bjo.2021.960. Ch03 n.3
- Patti L. Ephraim, Stephen T. Wegener, Ellen J. MacKenzie, Timothy R. Dillingham, and Liliana E. Pezzin. "Phantom Pain, Residual Limb Pain, and Back Pain in Amputees: Results of a National Survey." *Archives of Physical Medicine and Rehabilitation* 86, no. 10 (2005): 1910–1919. DOI 10.1016/j.apmr.2005.03.031. Ch07 n.1
- Nicholas Epley, Adam Waytz, and John T. Cacioppo. "On Seeing Human: A Three-Factor Theory of Anthropomorphism." *Psychological Review* 114, no. 4 (2007): 864–886. DOI 10.1037/0033-295X.114.4.864. Ch25 n.4

- Gareth Evans. *The Varieties of Reference*, ed. John McDowell (Oxford: Clarendon Press, 1982). Ch12 n.5, Ch12 n.6, Ch12 n.9, Ch12 n.10, Ch12 n.13, Ch14 n.8, Ch15 n.11, Ch16 n.16
- Charles Fernyhough. *The Voices Within: The History and Science of How We Talk to Ourselves* (New York: Basic Books, 2016). Ch17 n.3
- Richard P. Feynman. "Cargo Cult Science." Caltech commencement address, 1974. Preface epigraph
- Dominic H. ffytche. "Visual Hallucinations and the Charles Bonnet Syndrome." *Current Psychiatry Reports* 7, no. 3 (2005): 168–179. DOI 10.1007/s11920-005-0050-3. Ch03 n.3
- Kit Fine. "Guide to Ground." In *Metaphysical Grounding: Understanding the Structure of Reality*, ed. Fabrice Correia and Benjamin Schnieder (Cambridge: Cambridge University Press, 2012), 37–80. DOI 10.1017/CBO9781139149136.002. Ch06 n.21b
- J. R. Firth. "A Synopsis of Linguistic Theory, 1930–1955." In *Studies in Linguistic Analysis* (Oxford: Basil Blackwell, 1957), 1–32. Ch13 n.8, Ch23 n.5
- John Martin Fischer and Mark Ravizza. *Responsibility and Control: A Theory of Moral Responsibility* (Cambridge: Cambridge University Press, 1998). Ch18 n.5
- William Fish. *Perception, Hallucination, and Illusion* (Oxford: Oxford University Press, 2009). Argument from Hallucination n.7
- Luciano Floridi. "AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models." *Philosophy & Technology* 36, no. 1 (2023), article 15. Ch21 n.7
- Jerry A. Fodor. *A Theory of Content and Other Essays* (Cambridge, MA: MIT Press, 1990). Ch13 n.6, Ch15 n.8, App. B n.49
- Jerry A. Fodor. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (Cambridge, MA: MIT Press, 1987). Ch20 n.12, App. B n.49
- Jerry A. Fodor. *The Language of Thought* (New York: Crowell, 1975). Ch17 n.1
- John Foster. *The Nature of Perception* (Oxford: Oxford University Press, 2000). DOI 10.1093/0198237693.001.0001.
- Harry G. Frankfurt. "Alternate Possibilities and Moral Responsibility." *The Journal of Philosophy* 66, no. 23 (1969): 829–839. Ch18 n.5
- Harry G. Frankfurt. "Freedom of the Will and the Concept of a Person." *The Journal of Philosophy* 68, no. 1 (1971): 5–20. DOI 10.2307/2024717. Ch18 n.5
- Keith Frankish. "Illusionism as a Theory of Consciousness." *Journal of Consciousness Studies* 23, nos. 11–12 (2016): 11–39. App. B n.53
- Gottlob Frege. "Über Sinn und Bedeutung." *Zeitschrift für Philosophie und philosophische Kritik*, n.s. 100, no. 1 (1892): 25–50; trans. Max Black as "On Sense and Reference," in *Translations from the Philosophical Writings of Gottlob Frege*, ed. Peter Geach and Max Black (Oxford: Blackwell, 1952), 56–78. Ch16 n.2

- Karl Friston. "The Free-Energy Principle: A Unified Brain Theory?" *Nature Reviews Neuroscience* 11, no. 2 (2010): 127–138. Ch22 n.16
- Shaun Gallagher and Dan Zahavi. *The Phenomenological Mind*, 2nd ed. (London: Routledge, 2012). Ch19 n.4
- James Genone. "Appearance and Illusion." *Mind* 123, no. 490 (2014): 339–376. Argument from Illusion n.13
- James Genone. "Recent Work on Naïve Realism." *American Philosophical Quarterly* 53, no. 1 (2016): 1–25. Ch04 n.2, Ch06 n.17, Argument from Illusion n.13
- Edmund L. Gettier. "Is Justified True Belief Knowledge?" *Analysis* 23, no. 6 (1963): 121–123. Ch03 n.8
- Peter Godfrey-Smith. "Mental Representation, Naturalism, and Teleosemantics." In *Teleosemantics: New Philosophical Essays*, ed. Graham Macdonald and David Papineau (Oxford: Oxford University Press, 2006), 42–68. doi:10.1093/oso/9780199270262.003.0003. Ch14 n.10, Ch15 n.2, App. B n.47
- Peter Godfrey-Smith. "Mind, Matter, and Metabolism." *Journal of Philosophy* 113, no. 10 (2016): 481–506. DOI 10.5840/jphil20161131034. Ch24 n.6
- Peter Godfrey-Smith. "Triviality Arguments against Functionalism." *Philosophical Studies* 145, no. 2 (2009): 273–295. DOI 10.1007/s11098-008-9231-3. Ch22 n.4
- Peter Godfrey-Smith. *Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness* (New York: Farrar, Straus and Giroux, 2016). Ch24 n.6
- Philip Goff. "A Posteriori Physicalists Get Our Phenomenal Concepts Wrong." *Australasian Journal of Philosophy* 89, no. 2 (2011): 191–209. DOI 10.1080/00048401003649617. Ch10 n.3, Ch11 n.19, App. B n.37
- Philip Goff. "How Exactly Does Panpsychism Help Explain Consciousness?" *Journal of Consciousness Studies* 31, nos. 3–4 (2024): 56–82. DOI 10.53765/20512201.31.3.056. Ch10 n.2, Ch10 n.9, App. B n.32
- Philip Goff. "Real Acquaintance and Physicalism." In *Phenomenal Qualities: Sense, Perception, and Consciousness*, ed. Paul Coates and Sam Coleman (Oxford University Press, 2015), 121–143. DOI 10.1093/acprof:oso/9780198712718.003.0005. Ch10 n.11
- Philip Goff. *Consciousness and Fundamental Reality* (New York: Oxford University Press, 2017). Ch10 n.4
- Philip Goff. *Galileo's Error: Foundations for a New Science of Consciousness* (New York: Pantheon, 2019). Ch10 n.4
- I. J. Good. "Rational Decisions." *Journal of the Royal Statistical Society*, Series B, 14, no. 1 (1952): 107–114. App. B n.51a

- Nikola Grahek. *Feeling Pain and Being in Pain*, 2nd ed. (Cambridge, MA: MIT Press, 2007). DOI 10.7551/mitpress/2978.001.0001. Ch07 n.6
- Heather M. Gray, Kurt Gray, and Daniel M. Wegner. "Dimensions of Mind Perception." *Science* 315, no. 5812 (2007): 619. DOI 10.1126/science.1134475. Ch25 n.4
- Mitchell Green. "Speech Acts." *Stanford Encyclopedia of Philosophy*, substantive revision September 24, 2020. https://plato.stanford.edu/entries/speech-acts/. Ch16 n.1
- Daniel Gregory and Peter Langland-Hassan. "Inner Speech." *The Stanford Encyclopedia of Philosophy* (Fall 2023 edition).
- H. P. Grice. "Logic and Conversation." In *Syntax and Semantics*, vol. 3, *Speech Acts*, ed. Peter Cole and Jerry L. Morgan (New York: Academic Press, 1975), 41–58. Ch16 n.10
- H. P. Grice. "Meaning." *The Philosophical Review* 66, no. 3 (1957): 377–388. DOI 10.2307/2182440.
- Trevor Griffith and Adrian Kind. "Pain Asymbolia Is Not Pain." *Philosophy of Science* 91, no. 3 (2024): 561–578. DOI 10.1017/psa.2023.167. Ch07 n.6
- Jumbly Grindrod. "Large Language Models and Linguistic Intentionality." *Synthese* 204:71 (2024). Ch14 n.8, Ch22 n.6, Ch23 n.7
- Daya Guo et al. "DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning." *Nature* 645 (2025): 633–638. DOI 10.1038/s41586-025-09422-z. Ch23 n.10
- Wes Gurnee and Max Tegmark. "Language Models Represent Space and Time." *International Conference on Learning Representations* (2024). arXiv:2310.02207. Ch13 n.9
- Wes Gurnee et al. "Verbalizable Representations Form a Global Workspace in Language Models." *Transformer Circuits Thread*, July 6, 2026; arXiv:2607.15495v1, July 16, 2026. Ch24 n.12, Coda n.4, App. B n.68
- Stewart Elliott Guthrie. *Faces in the Clouds: A New Theory of Religion* (New York: Oxford University Press, 1993). Ch25 n.6
- P. M. S. Hacker. *Wittgenstein: Meaning and Mind* (Oxford: Blackwell, 1990). Ch19 n.2
- C. L. Hardin. *Color for Philosophers: Unweaving the Rainbow* (Indianapolis: Hackett, 1988). App. B n.22
- Gilbert Harman. "The Intrinsic Quality of Experience." *Philosophical Perspectives* 4 (1990): 31–52. DOI 10.2307/2214186. Ch04 n.1
- Gilbert Harman. "Wide Functionalism." In *Reasoning, Meaning, and Mind* (New York: Oxford University Press, 1999), 235–243. Ch20 n.10
- Stevan Harnad. "The Symbol Grounding Problem." *Physica D: Nonlinear Phenomena* 42, nos. 1–3 (1990): 335–346. Ch13 n.2, Ch22 n.19, Ch23 n.3, Ch24 n.11
- Zellig S. Harris. "Distributional Structure." *Word* 10, nos. 2–3 (1954): 146–162. Ch13 n.8

- Gary Hatfield. "The Cognitive Faculties." In *The Cambridge History of Seventeenth-Century Philosophy*, ed. Daniel Garber and Michael Ayers, 951–1002 (Cambridge: Cambridge University Press, 1998). DOI 10.1017/CHOL9780521572330.002. Ch02 n.2
- John Haugeland. "Understanding Natural Language." *The Journal of Philosophy* 76, no. 11 (1979): 619–632. App. B.12 (in text)
- John Haugeland. *Having Thought: Essays in the Metaphysics of Mind* (Cambridge, MA: Harvard University Press, 1998). App. B.12 (in text)
- Vladimír Havlík. "Meaning and Understanding in Large Language Models." *Synthese* 205, no. 1 (2025): article 9. DOI 10.1007/s11229-024-04878-4. Ch14 n.12, Ch16 n.20, Ch22 n.17
- John Hawthorne and Karson Kovakovich. "Disjunctivism." *Aristotelian Society Supplementary Volume* 80, no. 1 (2006): 145–183. doi:10.1111/j.1467-8349.2006.00141.x. Argument from Hallucination n.12
- Martin Hägglund. *This Life: Secular Faith and Spiritual Freedom* (New York: Pantheon, 2019). Ch24 n.9
- Martin Heidegger. *Being and Time* (1927). Ch20 n.22
- Fritz Heider and Marianne Simmel. "An Experimental Study of Apparent Behavior." *The American Journal of Psychology* 57, no. 2 (1944): 243–259. Ch25 n.3
- John Heil. *The Universe As We Find It* (Oxford: Oxford University Press, 2012). Ch10 n.11
- David R. Hilbert. *Color and Color Perception: A Study in Anthropocentric Realism* (Stanford: CSLI Publications, 1987). App. B n.25
- Christopher S. Hill. *Sensations: A Defense of Type Materialism* (Cambridge: Cambridge University Press, 1991). Ch09 n.5
- Kashmir Hill. "They Asked an A.I. Chatbot Questions. The Answers Sent Them Spiraling." *The New York Times*, June 2025. Preface n.3
- Geoffrey Hinton. "The Forward-Forward Algorithm: Some Preliminary Investigations." arXiv:2212.13345 (2022). Ch24 n.10
- J. M. Hinton. "Visual Experiences." *Mind* 76, no. 302 (1967): 217–227. Ch05 n.7
- J. M. Hinton. *Experiences: An Inquiry into Some Ambiguities* (Oxford: Clarendon Press, 1973). Ch05 n.7
- Jakob Hohwy. "The Self-Evidencing Brain." *Noûs* 50, no. 2 (2016): 259–285. doi:10.1111/nous.12062. Ch22 n.16
- Jakob Hohwy. *The Predictive Mind* (Oxford: Oxford University Press, 2013). Ch02 n.7
- Robert J. Howell. "The Knowledge Argument and the Implications of Phenomenal Knowledge." *Philosophy Compass* 6, no. 7 (2011): 459–468. DOI 10.1111/j.1747-9991.2011.00408.x. Ch08 n.14

- Robert J. Howell. *Consciousness and the Limits of Objectivity: The Case for Subjective Physicalism* (Oxford University Press, 2013). DOI 10.1093/acprof:oso/9780199654666.001.0001. Ch08 n.14
- David Hume. *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects*. 3 vols. (London: John Noon, 1739; Thomas Longman, 1740). Ch17 n.7, Ch02 (epigraph), Ch17 (epigraph)
- Paul Humphreys. "Why Propensities Cannot Be Probabilities." *Philosophical Review* 94, no. 4 (1985): 557–570. App. B n.51a
- Russell T. Hurlburt, Christopher L. Heavey, and Jason M. Kelsey. "Toward a Phenomenology of Inner Speaking." *Consciousness and Cognition* 22, no. 4 (2013): 1477–1494. Ch17 n.3
- Edmund Husserl. *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy, First Book*, trans. F. Kersten (The Hague: Martinus Nijhoff, 1982); earlier trans. W. R. Boyce Gibson as *Ideas: General Introduction to Pure Phenomenology* (London: George Allen & Unwin, 1931). Intro n.1; epigraph
- Edmund Husserl. *Logical Investigations*, trans. J. N. Findlay, ed. with an introduction by Dermot Moran, 2 vols. (London: Routledge, 2001). Ch01 n.1
- Daniel D. Hutto and Erik Myin. *Radicalizing Enactivism: Basic Minds without Content* (Cambridge, MA: MIT Press, 2013). Ch01 n.1, Ch15 n.10
- Frank Jackson. "Epiphenomenal Qualia." *The Philosophical Quarterly* 32, no. 127 (1982): 127–136. Ch06 n.12, Ch08 n.1, Ch08 n.2, Ch08 n.6, Ch09 n.8, Ch11 n.7
- Frank Jackson. "Mind and Illusion." In *Minds and Persons*, ed. Anthony O'Hear, *Royal Institute of Philosophy Supplement* 53 (Cambridge: Cambridge University Press, 2003), 251–272. Ch06 n.14, Ch08 nn.5–6, Ch11 n.7
- Frank Jackson. "What Mary Didn't Know." *The Journal of Philosophy* 83, no. 5 (1986): 291–295.
- Frank Jackson. *Perception: A Representative Theory* (Cambridge: Cambridge University Press, 1977). Ch03 n.4
- Pierre Jacob. "Intentionality." *Stanford Encyclopedia of Philosophy*, substantive revision February 7, 2023. https://plato.stanford.edu/entries/intentionality/.
- Hilla Jacobson. "Killing the Messenger: Representationalism and the Painfulness of Pain." *Philosophical Quarterly* 63, no. 252 (2013): 509–519. doi:10.1111/1467-9213.12029. Ch07 n.5
- William James. "The Dilemma of Determinism" (1884). In *The Will to Believe, and Other Essays in Popular Philosophy* (New York: Longmans, Green, 1912), 145–183. First collected 1897. Ch18 n.9
- William James. *The Principles of Psychology*, vol. 1 (New York: Henry Holt and Company, 1890). Ch10 n.8
- E. T. Jaynes. "Probability Theory as Logic." In *Maximum Entropy and Bayesian Methods*, ed. Paul F. Fougère (Dordrecht: Kluwer, 1990), 1–16. DOI 10.1007/978-94-009-0683-9_1. App. B n.51a

- Julian Jaynes. *The Origin of Consciousness in the Breakdown of the Bicameral Mind* (Boston: Houghton Mifflin, 1976). Ch17 n.8
- Mark Johnston. "The Obscure Object of Hallucination." *Philosophical Studies* 120, nos. 1–3 (2004): 113–183. Argument from Hallucination n.7
- Hans Jonas. *The Phenomenon of Life: Toward a Philosophical Biology* (New York: Harper & Row, 1966). Ch24 nn.5, 9, App. B n.39
- Cameron R. Jones and Benjamin K. Bergen. "Large Language Models Pass the Turing Test." arXiv:2503.23674 (2025). DOI 10.48550/arXiv.2503.23674. Ch24 n.2
- Kevin Jung. "Augustine, AI, and the Two Models of Language." *Journal of Religious Ethics* 53, no. 2 (2025): 217–238. doi:10.1111/jore.70004. App. B n.4
- François Kammerer. "Can You Believe It? Illusionism and the Illusion Meta-Problem." *Philosophical Psychology* 31, no. 1 (2018): 44–67. DOI 10.1080/09515089.2017.1388361.
- François Kammerer. "The Hardest Aspect of the Illusion Problem — and How to Solve It." *Journal of Consciousness Studies* 23, nos. 11–12 (2016): 124–139. App. B n.56
- Ryota Kanai and Shuqin Ma. "Intrinsic Computational Functionalism and Simulated Consciousness." arXiv:2606.15348 (2026). Ch22 n.5, Ch24 n.10
- Robert Kane. *The Significance of Free Will* (New York: Oxford University Press, 1996). App. B n.65
- Immanuel Kant. *Critique of Practical Reason*, trans. Werner S. Pluhar, with an introduction by Stephen Engstrom (Indianapolis: Hackett, 2002 [1788]). Ch18 n.9
- Immanuel Kant. *Critique of Pure Reason* (1781/1787). App. B n.64
- Webb Keane. "From Talking Tools to Metahumans: Social Interaction, Semiotic Skill, and the Authority of AI Chatbots." *Journal of the Royal Anthropological Institute* 32, no. 3 (2026): 1022–1039. doi:10.1111/1467-9655.70133. Ch25 n.4
- Brian W. Kernighan and Dennis M. Ritchie. *The C Programming Language* (Englewood Cliffs, NJ: Prentice-Hall, 1978). Ch20 n.2
- Brian W. Kernighan. *Programming in C: A Tutorial* (Murray Hill, NJ: Bell Laboratories, 1974). Ch20 n.2
- Jaegwon Kim. *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation* (Cambridge, MA: MIT Press, 1998). Ch09 n.7, Ch18 n.6
- Matthew Kinakin. "Why Intentionalists Can't Take Painkillers." *Philosophical Studies* 183, no. 6 (2026): 1905–1929. DOI 10.1007/s11098-026-02523-z. Ch07 n.5
- Amy Kind. "What's So Transparent about Transparency?" *Philosophical Studies* 115, no. 3 (2003): 225–244. Ch04 n.4, Ch06 n.3, Ch06 n.9, Ch11 n.19, Transparency n.4
- G. S. Kirk, J. E. Raven, and M. Schofield. *The Presocratic Philosophers: A Critical History with a Selection of Texts.* 2nd ed. (Cambridge: Cambridge University Press, 1983). Ch25 n.8

- Colin Klein. “An Imperative Theory of Pain.” *Journal of Philosophy* 104, no. 10 (2007): 517–532. doi:10.5840/jphil2007104104. Ch07 n.5
- Colin Klein. *What the Body Commands: The Imperative Theory of Pain* (Cambridge, MA: MIT Press, 2015).
- Kepa Korta and John Perry. “Pragmatics.” *Stanford Encyclopedia of Philosophy*, substantive revision May 28, 2024. https://plato.stanford.edu/entries/pragmatics/. Ch16 n.1
- Alfred Korzybski. *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics*, 5th ed. (Brooklyn, NY: Institute of General Semantics, 1994; second printing, 2000). Ch22 (epigraph)
- Alex Kozulin. “Vygotsky in Context.” Introduction to Lev Vygotsky, *Thought and Language*, translation newly revised and edited by Alex Kozulin (Cambridge, MA: MIT Press, 1986). App. B n.1
- Uriah Kriegel, ed. *The Routledge Handbook of Franz Brentano and the Brentano School* (London: Routledge, 2017). Ch01 n.1, Ch13 n.4, App. B n.28
- Uriah Kriegel. *Subjective Consciousness: A Self-Representational Theory* (Oxford University Press, 2009). Ch06 n.7b, App. B n.59
- Saul A. Kripke. *Naming and Necessity* (Cambridge, MA: Harvard University Press, 1980). Ch08 n.12, Ch11 n.5, Ch16 n.13, App. B n.12
- Saul Kripke. *Wittgenstein on Rules and Private Language: An Elementary Exposition* (Cambridge, MA: Harvard University Press, 1982). Ch17 n.5
- Frederick W. Kroon. “Causal Descriptivism.” *Australasian Journal of Philosophy* 65, no. 1 (1987): 1–17. DOI 10.1080/00048408712342731. Ch12 n.9
- Ray Kurzweil. *How to Create a Mind: The Secret of Human Thought Revealed* (New York: Viking, 2012). Ch24 n.8
- Ray Kurzweil. *The Age of Spiritual Machines: When Computers Exceed Human Intelligence* (New York: Viking, 1999). Coda epigraph
- Ray Kurzweil. *The Singularity Is Near: When Humans Transcend Biology* (New York: Viking, 2005). Ch24 n.8, Coda n.1
- Nathan Lambert et al. “Tulu 3: Pushing Frontiers in Open Language Model Post-Training.” arXiv:2411.15124 (2024). Ch23 n.10
- Victor A. F. Lamme. “Towards a True Neural Stance on Consciousness.” *Trends in Cognitive Sciences* 10, no. 11 (2006): 494–501. doi:10.1016/j.tics.2006.09.001. Ch24 n.15, App. B n.76
- Hakwan Lau and David Rosenthal. “Empirical Support for Higher-Order Theories of Conscious Awareness.” *Trends in Cognitive Sciences* 15, no. 8 (2011): 365–373. Ch06 n.7b, App. B n.57
- Shane Legg and Marcus Hutter. “Universal Intelligence: A Definition of Machine Intelligence.” arXiv:0712.3329v1 (2007). Intro n.2

- Gottfried Wilhelm Leibniz. “Discourse on Metaphysics” (1686). In *Philosophical Essays*, ed. and trans. Roger Ariew and Daniel Garber (Hackett, 1989). Ch12 n.8
- Alessandro Lenci. “Distributional Models of Word Meaning.” *Annual Review of Linguistics* 4 (2018): 151–171. Ch23 n.5
- Alexander Lerchner. “The Abstraction Fallacy: Why AI Can Simulate But Not Instantiate Consciousness.” Manuscript, March 19, 2026. PhilArchive, LERTAF. Coda n.6
- J. Y. Lettvin, H. R. Maturana, W. S. McCulloch, and W. H. Pitts. “What the Frog’s Eye Tells the Frog’s Brain.” *Proceedings of the IRE* 47, no. 11 (1959): 1940–1951. DOI 10.1109/JRPROC.1959.287207. Ch15 epigraph
- Michael Levin. “Technological Approach to Mind Everywhere: An Experimentally-Grounded Framework for Understanding Diverse Bodies and Minds.” *Frontiers in Systems Neuroscience* 16 (2022): 768201. DOI 10.3389/fnsys.2022.768201. Ch15 n.9d
- Joseph Levine. “Materialism and Qualia: The Explanatory Gap.” *Pacific Philosophical Quarterly* 64, no. 4 (1983): 354–361. doi:10.1111/j.1468-0114.1983.tb00207.x. Ch09 n.2, Ch11 n.1, App. B n.10, Ch11 epigraph
- Joseph Levine. *Purple Haze: The Puzzle of Consciousness* (New York: Oxford University Press, 2001). Ch11 n.1
- Neil Levy. *Hard Luck: How Luck Undermines Free Will and Moral Responsibility* (Oxford University Press, 2011). DOI 10.1093/acprof:oso/9780199601387.001.0001. Ch18 n.3
- David K. Lewis. “An Argument for the Identity Theory.” *The Journal of Philosophy* 63, no. 1 (1966): 17–25. Ch20 n.5
- David Lewis. “A Subjectivist’s Guide to Objective Chance.” In *Studies in Inductive Logic and Probability*, vol. 2, ed. Richard C. Jeffrey (Berkeley: University of California Press, 1980), 263–293. App. B n.51a
- David Lewis. “Are We Free to Break the Laws?” *Theoria* 47, no. 3 (1981): 113–121. App. B n.65
- David Lewis. “Psychophysical and Theoretical Identifications.” *Australasian Journal of Philosophy* 50, no. 3 (1972): 249–258. DOI 10.1080/00048407212341301. Ch20 n.5
- David Lewis. “Putnam’s Paradox.” *Australasian Journal of Philosophy* 62, no. 3 (1984): 221–236. doi:10.1080/00048408412340013. Ch12 n.9, Ch22 n.20
- David Lewis. “Truth in Fiction.” *American Philosophical Quarterly* 15, no. 1 (1978): 37–46. Ch17 n.10
- David Lewis. “What Experience Teaches.” *Proceedings of the Russellian Society* 13 (1988): 29–57; reprinted in *Mind and Cognition: A Reader*, ed. William G. Lycan (Oxford: Blackwell, 1990), 499–519. Ch08 n.10, App. B n.16
- Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. “Emergent World Representations: Exploring a Sequence Model Trained on a

Synthetic Task." *International Conference on Learning Representations* (2023). arXiv:2210.13382. Ch12 n.12, Ch13 n.9, Ch22 n.18

- Benjamin Libet. "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action." *Behavioral and Brain Sciences* 8, no. 4 (1985): 529–539. App. B n.63
- Brian Loar. "Phenomenal States." *Philosophical Perspectives* 4 (1990): 81–108; revised in *The Nature of Consciousness: Philosophical Debates*, ed. Ned Block, Owen Flanagan, and Güven Güzeldere (Cambridge, MA: MIT Press, 1997), 597–616. Ch08 n.14, Ch09 n.5, Ch11 nn.6, 8, 12, App. B nn.16, 17, 36, 40
- John Locke. *An Essay Concerning Human Understanding*, ed. Peter H. Nidditch (Oxford: Clarendon Press, 1975). Ch02 n.3, Ch02 n.10, App. B n.23
- Jonathan Bennett. *Locke, Berkeley, Hume: Central Themes* (Oxford: Clarendon Press, 1971). Ch02 nn.3, 4, 5
- Heather Logue. "What Should the Naïve Realist Say about Total Hallucinations?" *Philosophical Perspectives* 26, no. 1 (2012): 173–199. Argument from Hallucination n.7
- Peter Ludlow, Yujin Nagasawa, and Daniel Stoljar, eds. *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument* (Cambridge, MA: MIT Press, 2004). Ch06 n.12, Ch08 n.3
- Peter Ludlow. "Descriptions." *Stanford Encyclopedia of Philosophy*, substantive revision September 21, 2022. https://plato.stanford.edu/entries/descriptions/. Ch16 n.3
- William G. Lycan. "Representational Theories of Consciousness." *Stanford Encyclopedia of Philosophy*, substantive revision August 12, 2019 (Winter 2019 archive). Inverted Earth n.17
- William G. Lycan. "The Intentionality of Smell." *Frontiers in Psychology* 5 (2014): 436. doi:10.3389/fpsyg.2014.00436. Ch07 n.8
- William G. Lycan. *Consciousness and Experience* (Cambridge, MA: MIT Press, 1996). Ch07 n.8, Inverted Earth nn.7, 17
- William G. Lycan. *Consciousness* (Cambridge, MA: MIT Press, 1987). App. B n.42
- Shuqin Ma and Ryota Kanai. "Intrinsic Computational Functionalism: From Observer-Relative Maps to Observer-Independent Structures." arXiv:2606.06424 (2026). Ch22 n.5, Ch24 n.10
- Fintan Mallory. "Teleosemantics for Neural Word Embeddings." *Mind & Language*, published online August 9, 2026, article mila.70037. DOI 10.1111/mila.70037. Ch14 n.11, Ch15 n.9b
- Jeff Malpas. "Donald Davidson." *Stanford Encyclopedia of Philosophy*, substantive revision April 28, 2023. https://plato.stanford.edu/entries/davidson/. Ch16 n.6
- Matthew Mandelkern and Tal Linzen. "Do Language Models' Words Refer?" *Computational Linguistics* 50, no. 3 (2024): 1191–1200. DOI 10.1162/coli_a_00522. Ch12 n.12, Ch16 n.18, Ch23 n.12, Ch24 n.11

- Samuel Marks and Max Tegmark. "The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets." arXiv:2310.06824 (2023). Ch23 n.10b
- M. G. F. Martin. "On Being Alienated." In *Perceptual Experience*, ed. Tamar Szabó Gendler and John Hawthorne (Oxford: Oxford University Press, 2006), 354–410. doi:10.1093/acprof:oso/9780199289769.003.0011. Ch05 n.7, Ch06 n.18, Ch23 n.14, Argument from Hallucination n.7
- M. G. F. Martin. "The Limits of Self-Awareness." *Philosophical Studies* 120 (2004): 37–89. Ch03 n.7, Ch05 n.7, Argument from Hallucination n.3, Argument from Hallucination n.7, Argument from Hallucination n.10
- M. G. F. Martin. "The Transparency of Experience." *Mind & Language* 17, no. 4 (2002): 376–425. DOI 10.1111/1468-0017.00205. Ch05 n.7, Ch06 n.18, Ch23 n.14, App. B n.74, Argument from Hallucination nn.7, 10
- Manolo Martínez. "Imperative Content and the Painfulness of Pain." *Phenomenology and the Cognitive Sciences* 10, no. 1 (2011): 67–90. doi:10.1007/s11097-010-9172-0. Ch07 n.5
- Humberto R. Maturana and Francisco J. Varela. *Autopoiesis and Cognition: The Realization of the Living* (Dordrecht: D. Reidel, 1980). Ch24 n.5
- John McCarthy and Patrick J. Hayes. "Some Philosophical Problems from the Standpoint of Artificial Intelligence." In *Machine Intelligence 4*, ed. B. Meltzer and D. Michie (Edinburgh: Edinburgh University Press, 1969), 463–502. Ch20 n.22
- John McDowell. "Criteria, Defeasibility, and Knowledge." *Proceedings of the British Academy* 68 (1982). Ch03 n.5, Ch05 n.7, Ch19 n.2, Argument from Hallucination n.2
- John McDowell. *Mind and World* (Cambridge, MA: Harvard University Press, 1994). Ch03 n.6, Ch06 n.20, App. B n.1, Argument from Hallucination nn.2, 10
- Colin McGinn. "Another Look at Color." *The Journal of Philosophy* 93, no. 11 (1996): 537–553. DOI 10.2307/2941048. App. B n.24
- Michael McKenna. *Conversation and Responsibility* (New York: Oxford University Press, 2012). Ch18 n.5
- Alexius Meinong. "The Theory of Objects." Trans. Isaac Levi, D. B. Terrell, and Roderick M. Chisholm. In *Realism and the Background of Phenomenology*, ed. Roderick M. Chisholm (Glencoe, IL: Free Press, 1960), 76–117. App. B n.29
- Alfred R. Mele. *Free Will and Luck* (New York: Oxford University Press, 2006). Ch18 n.10
- R. Melzack. "Phantom Limbs and the Concept of a Neuromatrix." *Trends in Neurosciences* 13, no. 3 (1990): 88–92. Ch07 n.1
- Angela Mendelovici and David Bourget. "Consciousness and Intentionality." In *The Oxford Handbook of the Philosophy of Consciousness*, ed. Uriah Kriegel (Oxford University Press, 2020), 559–585. doi:10.1093/oxfordhb/9780198749677.013.26. App. B n.52

- Angela Mendelovici. "Intentionalism about Moods." *Thought: A Journal of Philosophy* 2, no. 2 (2013): 126–136. DOI 10.1002/tht3.81. Ch07 nn.8, 10, App. B n.27
- Angela Mendelovici. "Pure Intentionalism about Moods and Emotions." In *Current Controversies in Philosophy of Mind*, ed. Uriah Kriegel (Routledge, 2014), 135–157. Ch07 n.10, App. B n.27
- Angela Mendelovici. "Reliable Misrepresentation and Tracking Theories of Mental Representation." *Philosophical Studies* 165, no. 2 (2013): 421–443. DOI 10.1007/s11098-012-9966-8. Ch13 n.1, App. B n.44, App. B n.52
- Maurice Merleau-Ponty. *Phenomenology of Perception*, trans. Colin Smith (London: Routledge & Kegan Paul, 1962 [1945]). Ch15 n.11, Ch20 n.22
- P. F. Strawson. *Individuals: An Essay in Descriptive Metaphysics* (London: Methuen, 1959). Ch12 n.4, Ch12 n.7, Ch12 epigraph
- Julian Michael. "To Dissect an Octopus: Making Sense of the Form/Meaning Debate." July 23, 2020. https://julianmichael.org/blog/2020/07/23/to-dissect-an-octopus.html. Ch23 n.1
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. "Distributed Representations of Words and Phrases and Their Compositionality." *Advances in Neural Information Processing Systems* 26 (2013): 3111–3119. App. B n.51a
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient Estimation of Word Representations in Vector Space." arXiv:1301.3781 (2013). Ch23 n.6
- Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. "Linguistic Regularities in Continuous Space Word Representations." In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Atlanta, GA: Association for Computational Linguistics, 2013), 746–751. Ch23 n.6
- Ruth Garrett Millikan. "Biosemantics." *The Journal of Philosophy* 86, no. 6 (1989): 281–297. Ch14 nn.6, 9, Ch15 nn.2, 4, 6–7
- Ruth Garrett Millikan. "On Swampkinds." *Mind & Language* 11, no. 1 (1996): 103–117. DOI 10.1111/j.1468-0017.1996.tb00035.x. Ch15 n.14
- Ruth Garrett Millikan. "What Has Natural Information to Do with Intentional Representation?" *Royal Institute of Philosophy Supplements* 49 (2001): 105–125. doi:10.1017/S135824610000713X. Ch14 n.12
- Ruth Garrett Millikan. *Language, Thought, and Other Biological Categories: New Foundations for Realism* (Cambridge, MA: MIT Press, 1984). Ch01 n.2, Ch06 n.19, Ch13 n.6, Ch14 n.6, Ch14 n.9, Ch14 n.10, Ch15 n.2, Ch15 n.4, Ch15 n.6, Ch15 n.7, Ch15 n.14, Ch22 n.6, Ch23 n.11, Ch23 n.15, Ch24 n.1, Ch24 n.4, App. B n.38, Inverted Earth n.13
- Ruth Garrett Millikan. *Varieties of Meaning: The 2002 Jean Nicod Lectures* (Cambridge, MA: MIT Press, 2004). Ch15 n.6
- Ruth Garrett Millikan. *White Queen Psychology and Other Essays for Alice* (Cambridge, MA: MIT Press, 1993). Ch15 n.2

- Kevin J. Mitchell. "The Origins of Meaning: From Pragmatic Control Signals to Semantic Representations." *Philosophy and the Mind Sciences* 7 (2026). DOI 10.33735/phimisci.2026.11674. Ch15 n.7
- Melanie Mitchell and David C. Krakauer. "The Debate over Understanding in AI's Large Language Models." *Proceedings of the National Academy of Sciences* 120, no. 13 (2023): e2215907120. doi:10.1073/pnas.2215907120. Ch24 n.14
- Dimitri Coelho Mollo and Raphaël Millière. "The Vector Grounding Problem." *Philosophy and the Mind Sciences* 7, no. 1 (2026): 1–28. DOI 10.33735/phimisci.2026.12307. Ch16 n.18
- Maël Montévil and Matteo Mossio. "Biological Organisation as Closure of Constraints." *Journal of Theoretical Biology* 372 (2015): 179–191. DOI 10.1016/j.jtbi.2015.02.029.
- G. E. Moore. "The Refutation of Idealism." *Mind* 12 (1903): 433–453. Ch04 n.4, Ch05 n.1, Ch06 n.3, Inverted Earth n.11, Transparency n.2
- Alvaro Moreno and Matteo Mossio. *Biological Autonomy: A Philosophical and Theoretical Enquiry* (Dordrecht: Springer, 2015). doi:10.1007/978-94-017-9837-2. Ch15 n.9a
- Thomas Nagel. "What Is It Like to Be a Bat?" *The Philosophical Review* 83, no. 4 (1974): 435–450. Ch09 n.1, Ch11 n.20
- Thomas Nagel. *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature Is Almost Certainly False* (New York: Oxford University Press, 2012). DOI 10.1093/acprof:oso/9780199919758.001.0001. Ch11 n.20
- Thomas Nagel. *The View from Nowhere* (New York: Oxford University Press, 1986). Ch09 n.1
- Karen Neander. "Swampman Meets Swampcow." *Mind & Language* 11, no. 1 (1996): 118–129. DOI 10.1111/j.1468-0017.1996.tb00036.x. Ch15 n.14
- Karen Neander. *A Mark of the Mental: In Defense of Informational Teleosemantics* (Cambridge, MA: MIT Press, 2017). Ch13 n.6, Ch14 n.5, Ch23 n.15, App. B n.46
- Johanne S. K. Nedergaard and Gary Lupyan. "Not Everybody Has an Inner Voice: Behavioral Consequences of Anendophasia." *Psychological Science* 35, no. 7 (2024): 780–797. DOI 10.1177/09567976241243004. Ch17 n.4
- Teodor Negru. "Intentionality and Background: Searle and Dreyfus against Classical AI Theory." *Filosofia Unisinos* 14, no. 1 (2013): 18–34. Ch20 n.24, Ch22 n.2
- Laurence Nemirow. "Physicalism and the Cognitive Role of Acquaintance." In *Mind and Cognition: A Reader*, ed. William G. Lycan (Oxford: Blackwell, 1990), 490–499. Ch08 n.10, App. B n.16
- Allen Newell and Herbert A. Simon. "Computer Science as Empirical Inquiry: Symbols and Search." *Communications of the ACM* 19, no. 3 (1976): 113–126. DOI 10.1145/360018.360022. Ch21 epigraph
- M. H. A. Newman. "Mr. Russell's 'Causal Theory of Perception'." *Mind* 37, no. 146 (1928): 137–148. doi:10.1093/mind/xxxvii.146.137. Ch22 n.18, Ch23 n.8

- Shaun Nichols and Joshua Knobe. “Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions.” *Noûs* 41, no. 4 (2007): 663–685. DOI 10.1111/j.1468-0068.2007.00666.x. App. B n.66
- Martine Nida-Rümelin and Donnchadh O Conaill. “Qualia: The Knowledge Argument.” *Stanford Encyclopedia of Philosophy*, substantive revision March 1, 2024. https://plato.stanford.edu/entries/qualia-knowledge/. Ch08 nn.3, 13
- Martine Nida-Rümelin. “Grasping Phenomenal Properties.” In *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, ed. Torin Alter and Sven Walter (New York: Oxford University Press, 2007), 307–338. DOI 10.1093/acprof:oso/9780195171655.003.0013. Ch08 n.13
- Alva Noë. *Action in Perception* (Cambridge, MA: MIT Press, 2004). Ch06 n.21d, App. B nn.23, 25
- Robert Nozick. *Philosophical Explanations* (Cambridge, MA: Belknap Press of Harvard University Press, 1981). Ch03 n.8
- Long Ouyang et al. “Training Language Models to Follow Instructions with Human Feedback.” arXiv:2203.02155 (2022). Ch23 n.10a
- Ísak Andri Ólafsson. “Artificial Competence.” *Episteme* (online 2026): 1–13. DOI 10.1017/epi.2026.10129. Ch24 n.14
- Sarah Palmer. “Love in the Time of AI: Woman Creates and ‘Marries’ AI-Powered Chatbot Boyfriend.” *Euronews*, June 7, 2023; updated June 8, 2023. https://www.euronews.com/next/2023/06/07/love-in-the-time-of-ai-woman-claims-she-married-a-chatbot-and-is-expecting-its-baby. Preface n.2
- Stefano Palminteri and Charley M. Wu. “Beyond Computational Equivalence: The Behavioral Inference Principle for Machine Consciousness.” *Neuroscience of Consciousness* 2026, no. 1 (2026): niag002. DOI 10.1093/nc/niag002. Ch24 n.12
- Ezequiel A. Di Paolo. “Autopoiesis, Adaptivity, Teleology, Agency.” *Phenomenology and the Cognitive Sciences* 4, no. 4 (2005): 429–452. DOI 10.1007/s11097-005-9002-y. Ch24 n.5
- Ezequiel Di Paolo. “Extended Life.” *Topoi* 28, no. 1 (2009): 9–21. DOI 10.1007/s11245-008-9042-3. Ch24 n.5
- David Papineau. “Mind the Gap.” *Philosophical Perspectives* 12 (1998): 373–388. Ch11 n.8
- David Papineau. “Physicalism, Consciousness and the Antipathetic Fallacy.” *Australasian Journal of Philosophy* 71, no. 2 (1993): 169–183. DOI 10.1080/00048409312345182. Ch11 n.8
- David Papineau. *Philosophical Naturalism* (Oxford: Blackwell, 1993). Ch11 n.8
- David Papineau. *Thinking About Consciousness* (Oxford: Oxford University Press, 2002). 0 — Preface n.4, Ch09 n.7, Ch11 n.8, App. B n.17, App. B n.40
- Derek Parfit. *Reasons and Persons* (Oxford: Oxford University Press, 1984). Ch17 n.7, Ch24 n.8

- Robert Pasnau. *Theories of Cognition in the Later Middle Ages* (Cambridge: Cambridge University Press, 1997). Ch06 n.3
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, 2014), 1532–1543. DOI 10.3115/v1/D14-1162.
- Roger Penrose. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics* (Oxford: Oxford University Press, 1989). App. B n.21
- Derk Pereboom and Michael McKenna. "Manipulation Arguments Against Compatibilism." In *The Oxford Handbook of Moral Responsibility*, edited by Dana Kay Nelkin and Derk Pereboom (Oxford University Press, 2022), 179–200. DOI: 10.1093/oxfordhb/9780190679309.013.38.
- Derk Pereboom. *Living Without Free Will* (Cambridge: Cambridge University Press, 2001). DOI 10.1017/CBO9780511498824. Ch18 n.2
- John Perry. "The Problem of the Essential Indexical." *Noûs* 13, no. 1 (1979): 3–21. Ch08 n.13
- John Perry. *Knowledge, Possibility, and Consciousness* (Cambridge, MA: MIT Press, 2001). Ch08 n.13
- Steven T. Piantadosi and Felix Hill. "Meaning without reference in large language models." arXiv:2208.02957 (2022).
- Gualtiero Piccinini. "Computational Modelling vs. Computational Explanation: Is Everything a Turing Machine, and Does It Matter to the Philosophy of Mind?" *Australasian Journal of Philosophy* 85, no. 1 (2007): 93–115. App. B n.20
- Gualtiero Piccinini. "Computationalism, the Church–Turing Thesis, and the Church–Turing Fallacy." *Synthese* 154, no. 1 (2007): 97–120. Ch22 n.10, App. B nn.8, 19
- Gualtiero Piccinini. *Physical Computation: A Mechanistic Account* (Oxford: Oxford University Press, 2015). App. B n.20
- Bryan Pickel and Zoltán Gendler Szabó. "Compositionality." *Stanford Encyclopedia of Philosophy*, substantive revision November 3, 2025. Ch16 n.4
- U. T. Place. "Is Consciousness a Brain Process?" *British Journal of Psychology* 47, no. 1 (1956): 44–50. DOI 10.1111/j.2044-8295.1956.tb00560.x. Ch06 n.21a
- Angela Mendelovici. *The Phenomenal Basis of Intentionality* (Oxford University Press, 2018). Ch06 n.21d, Ch07 n.10, Ch13 n.1, App. B n.52
- Daniel C. Dennett. *Brainchildren: Essays on Designing Minds* (MIT Press, 1998).
- Sam Harris. *Free Will* (Free Press, 2012).
- Karl R. Popper. "The Propensity Interpretation of Probability." *British Journal for the Philosophy of Science* 10, no. 37 (1959): 25–42. App. B n.51a

- H. H. Price. *Perception* (London: Methuen, 1932). Ch02 n.6, Ch03 n.2, Argument from Illusion nn.1, 10
- Duncan Pritchard. *Epistemic Luck* (Oxford: Oxford University Press, 2005). Ch03 n.8
- Hilary Putnam. "Philosophy and Our Mental Life." In *Mind, Language and Reality: Philosophical Papers*, vol. 2 (Cambridge: Cambridge University Press, 1975), 291–303. DOI 10.1017/CBO9780511625251.016. Ch20 n.4
- Hilary Putnam. "Psychological Predicates" (1967), reprinted as "The Nature of Mental States." In *Mind, Language and Reality: Philosophical Papers*, vol. 2 (Cambridge: Cambridge University Press, 1975), 429–440. DOI 10.1017/CBO9780511625251.023. Ch20 n.4
- Hilary Putnam. "The Meaning of 'Meaning'." In *Language, Mind, and Knowledge*, ed. Keith Gunderson, *Minnesota Studies in the Philosophy of Science* 7 (Minneapolis: University of Minnesota Press, 1975), 131–193. Ch14 n.9, Ch16 n.14, Ch20 n.8, App. B n.4
- Hilary Putnam. *Reason, Truth and History* (Cambridge: Cambridge University Press, 1981). DOI 10.1017/CBO9780511625398. Ch16 n.19
- Hilary Putnam. *Representation and Reality* (Cambridge, MA: MIT Press, 1988). Ch17 n.1, Ch20 nn.4, 6, 8, 9, 11, 12, Ch22 nn.1, 3
- Hilary Putnam. *Words and Life*, ed. James Conant (Cambridge, MA: Harvard University Press, 1994). Ch20 n.9
- V. S. Ramachandran and Sandra Blakeslee. *Phantoms in the Brain: Probing the Mysteries of the Human Mind* (New York: William Morrow, 1998).
- Matthew Ratcliffe. *Feelings of Being: Phenomenology, Psychiatry and the Sense of Reality* (Oxford: Oxford University Press, 2008). Ch07 n.9, App. B n.27
- Vasudevi Reddy. *How Infants Know Minds* (Harvard University Press, 2008). Ch19 n.4
- Byron Reeves and Clifford Nass. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places* (Stanford, CA: CSLI Publications; New York: Cambridge University Press, 1996). Ch25 n.5
- Flora Rheta Schreiber. *Sybil* (Chicago: Regnery, 1973).
- Thomas Reid. *Essays on the Intellectual Powers of Man* (Edinburgh: John Bell; London: G. G. J. & J. Robinson, 1785). Ch02 n.1
- Indrek Reiland. "The Unity of Perceptual Content." *The Philosophical Quarterly* 74, no. 3 (2024): 941–961. doi:10.1093/pq/pqad105. Transparency n.10, Argument from Hallucination n.5
- Michael Rescorla. "The Computational Theory of Mind." *The Stanford Encyclopedia of Philosophy*, first published October 16, 2015; substantive revision December 18, 2024. Ch22 n.5
- Howard Robinson. *Perception* (London: Routledge, 1994). Ch02 n.5, Ch03 nn.1, 4, Ch05 n.4, Argument from Illusion nn.3, 8, 11

- David M. Rosenthal. "Higher-Order Theories of Consciousness." In *The Oxford Handbook of Philosophy of Mind*, ed. Brian P. McLaughlin, Ansgar Beckermann, and Sven Walter (Oxford: Oxford University Press, 2009), 239–252. DOI 10.1093/oxfordhb/9780199262618.003.0014. Ch24 n.15
- David M. Rosenthal. *Consciousness and Mind* (Oxford: Clarendon Press, 2005). Ch06 n.7b
- David Rosenthal. "Exaggerated Reports: Reply to Block." *Analysis* 71, no. 3 (2011): 431–437. Ch06 n.7b, App. B n.58
- John Ruskin. *Modern Painters*. Vol. 3, part 4, *Of Many Things* (London: Smith, Elder, 1856). Ch25 n.10
- Bertrand Russell. "Knowledge by Acquaintance and Knowledge by Description." *Proceedings of the Aristotelian Society* 11 (1910–11): 108–128. Ch12 n.9
- Bertrand Russell. "On Denoting." *Mind*, n.s. 14, no. 56 (1905): 479–493. DOI 10.1093/mind/XIV.4.479.
- Bertrand Russell. *Human Knowledge: Its Scope and Limits* (London: George Allen & Unwin, 1948). Ch03 n.8
- Bertrand Russell. *The Analysis of Matter* (London: Kegan Paul, Trench, Trubner & Co., 1927). Ch10 n.1
- Bertrand Russell. *The Problems of Philosophy* (London: Williams and Norgate, 1912). Ch02 n.6, Ch04 n.2
- Gilbert Ryle. *The Concept of Mind* (London: Hutchinson, 1949). App. B n.64
- Robert M. Sapolsky. *Determined: A Science of Life Without Free Will* (New York: Penguin Press, 2023). Ch18 n.2
- Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. "Are Emergent Abilities of Large Language Models a Mirage?" *Advances in Neural Information Processing Systems* 36 (2023): 55565–55581. DOI 10.52202/075280-2425. Ch21 n.8
- Jonathan Schaffer. "Monism: The Priority of the Whole." *The Philosophical Review* 119, no. 1 (2010): 31–76. DOI 10.1215/00318108-2009-025. Ch10 n.9
- Roger C. Schank and Robert P. Abelson. *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures* (Hillsdale, NJ: Lawrence Erlbaum, 1977). Ch21 n.1
- Marya Schechtman. *The Constitution of Selves* (Ithaca: Cornell University Press, 1996). Ch17 n.7
- Matthias Scheutz. "When Physical Systems Realize Functions…" *Minds and Machines* 9, no. 2 (1999): 161–196. DOI 10.1023/A:1008364332419. Ch22 n.5
- Peter Schulte and Karen Neander. "Teleological Theories of Mental Content." *Stanford Encyclopedia of Philosophy*, substantive revision May 26, 2022. https://plato.stanford.edu/entries/content-teleological/. Ch15 n.8, Ch15 n.9, Ch15 n.9b, Ch15 n.14, App. B n.49

- Aaron Schurger, Jacobo D. Sitt, and Stanislas Dehaene. "An Accumulator Model for Spontaneous Neural Activity Prior to Self-Initiated Movement." *Proceedings of the National Academy of Sciences* 109, no. 42 (2012): E2904–E2913. doi:10.1073/pnas.1210467109. App. B n.63
- Wolfgang Schwarz. "Against Magnetism." *Australasian Journal of Philosophy* 92, no. 1 (2014): 17–36. DOI 10.1080/00048402.2013.765900. Ch22 n.20
- Eric Schwitzgebel. "AI Systems Must Not Confuse Users about Their Sentience or Moral Status." *Patterns* 4, no. 8 (2023): 100818. Coda n.10
- Eric Schwitzgebel. "Borderline Consciousness, When It's Neither Determinately True nor Determinately False That Experience Is Present." *Philosophical Studies* 180, no. 12 (2023): 3415–3439. Ch10 n.7
- William Seager. "Consciousness, Information and Panpsychism." *Journal of Consciousness Studies* 2, no. 3 (1995): 272–288. Ch10 n.9
- John R. Searle. "Is the Brain a Digital Computer?" *Proceedings and Addresses of the American Philosophical Association* 64, no. 3 (1990): 21–37. Ch13 n.7, Ch14 n.4, Ch20 n.25, Ch21 n.5, Ch22 n.2
- John R. Searle. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3, no. 3 (1980): 417–424. DOI 10.1017/S0140525X00005756. Ch06 n.19, Ch14 n.13, Ch20 n.25, Ch21 nn.1, 2, 4, 6, Ch22 n.8, Ch14 epigraph
- John R. Searle. "The Philosophy of Perception and the Bad Argument." In *Wirklichkeit oder Konstruktion? Sprachtheoretische und interdisziplinäre Aspekte einer brisanten Alternative*, ed. Ekkehard Felder and Andreas Gardt, 66–76 (Berlin: De Gruyter, 2018). DOI 10.1515/9783110563436-003. Ch06 n.16
- John R. Searle. *Intentionality: An Essay in the Philosophy of Mind* (Cambridge: Cambridge University Press, 1983). Ch01 n.2, Ch06 n.19, Ch13 n.3, App. B n.26
- John R. Searle. *Mind: A Brief Introduction* (New York: Oxford University Press, 2004). Ch11 n.15
- John R. Searle. *Minds, Brains and Science* (Cambridge, MA: Harvard University Press, 1984). Ch18 n.9
- John R. Searle. *Rationality in Action* (Cambridge, MA: MIT Press, 2001). Ch18 n.9
- John R. Searle. *Seeing Things as They Are: A Theory of Perception* (Oxford: Oxford University Press, 2015). Ch06 n.16
- John R. Searle. *Speech Acts: An Essay in the Philosophy of Language* (Cambridge: Cambridge University Press, 1969). Ch16 n.12
- John R. Searle. *The Mystery of Consciousness* (New York: New York Review Books, 1997). Ch14 n.4, Ch22 nn.2, 9, App. B n.6
- John R. Searle. *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992). Ch01 n.2, Ch11 nn.15, 16, Ch13 nn.3, 7, Ch21 n.5, Ch24 nn.3, 15, App. B n.30

- Jeff Sebo and Robert Long. “Moral Consideration for AI Systems by 2030.” *AI and Ethics* 5, no. 1 (2025): 591–606. DOI 10.1007/s43681-023-00379-1. App. B n.73
- Wilfrid Sellars. *Science, Perception and Reality* (London: Routledge & Kegan Paul, 1963). Ch17 n.6
- J. L. Austin. *Sense and Sensibilia*, reconstructed from the manuscript notes by G. J. Warnock (Oxford: Clarendon Press, 1962). Ch02 n.6, Ch03 nn.1, 2, 5, Ch05 n.6, Argument from Illusion nn.4, 11
- Anil Seth. *Being You: A New Science of Consciousness* (London: Faber & Faber, 2021). Ch02 n.7
- Oron Shagrir. “The Rise and Fall of Computational Functionalism.” In *Hilary Putnam*, ed. Yemima Ben-Menahem (Cambridge: Cambridge University Press, 2005), 220–250. doi:10.1017/CBO9780511614187.009. Ch22 n.3
- Murray Shanahan. “Talking about Large Language Models.” *Communications of the ACM* 67, no. 2 (2024): 68–79. doi:10.1145/3624724. Ch21 n.7
- Claude E. Shannon. “A Mathematical Theory of Communication.” *Bell System Technical Journal* 27 (1948): 379–423, 623–656. Ch13 n.5
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, et al. “Towards Understanding Sycophancy in Language Models.” In *The Twelfth International Conference on Learning Representations* (2024). arXiv:2310.13548. Ch23 n.10a
- Nicholas Shea. “Realist Representational Explanations of Agency Should Require Some Unity of Purpose.” *Philosophy and the Mind Sciences* 7, no. 1 (2026). DOI 10.33735/phimisci.2026.12098. Ch15 n.9b
- Nicholas Shea. *Representation in Cognitive Science* (Oxford: Oxford University Press, 2018). Ch01 n.2, Ch15 nn.7, 9, 9b, Ch24 nn.1, 4, App. B n.50
- Sydney Shoemaker. “Functionalism and Qualia.” *Philosophical Studies* 27, no. 5 (1975): 291–315. DOI 10.1007/BF01225748. App. B, inverted-spectrum discussion
- Sydney Shoemaker. “The Inverted Spectrum.” *The Journal of Philosophy* 79, no. 7 (1982): 357–381. Ch02 n.10, Ch06 n.14a
- Sydney Shoemaker. “Self-Knowledge and ‘Inner Sense’: Lecture III: The Phenomenal Character of Experience.” *Philosophy and Phenomenological Research* 54, no. 2 (1994): 291–314. DOI 10.2307/2108490. Ch04 n.2
- Susanna Siegel. *The Contents of Visual Experience* (Oxford: Oxford University Press, 2010). Argument from Illusion n.7
- Jonathan A. Simon. “Vagueness and Zombies: Why ‘Phenomenally Conscious’ Has No Borderline Cases.” *Philosophical Studies* 174, no. 8 (2017): 2105–2123. DOI 10.1007/s11098-016-0790-4. Ch10 n.6
- J. J. C. Smart. “Sensations and Brain Processes.” *The Philosophical Review* 68, no. 2 (1959): 141–156. Ch06 n.21a

- A. D. Smith. *The Problem of Perception* (Cambridge, MA: Harvard University Press, 2002). Ch05 n.4; Argument from Hallucination n.11; Argument from Illusion nn.12–13
- Brian Cantwell Smith. *The Promise of Artificial Intelligence: Reckoning and Judgment* (Cambridge, MA: MIT Press, 2019). Ch24 epigraph
- Kurt Smith. "Descartes' Theory of Ideas." *Stanford Encyclopedia of Philosophy*, substantive revision September 18, 2025. https://plato.stanford.edu/entries/descartes-ideas/. Ch02 n.1
- Paul Snowdon. "Perception, Vision and Causation." *Proceedings of the Aristotelian Society*, n.s., 81 (1980–81): 175–192. Ch03 n.6, Ch05 n.7, Argument from Hallucination n.4
- Paul Snowdon. "The Formulation of Disjunctivism: A Response to Fish." *Proceedings of the Aristotelian Society* 105, no. 1 (2005): 129–141.
- Ernest Sosa. *A Virtue Epistemology*. Vol. 1 of *Apt Belief and Reflective Knowledge* (Oxford: Clarendon Press, 2007). Ch03 n.8
- Matthew Soteriou. "The Disjunctive Theory of Perception." *Stanford Encyclopedia of Philosophy*, substantive revision July 28, 2026. Ch05 n.7, Ch06 n.17, Ch06 n.18, App. B n.74, Argument from Hallucination n.7
- Jeff Speaks. "Theories of Meaning." *Stanford Encyclopedia of Philosophy*, substantive revision July 31, 2024. https://plato.stanford.edu/entries/meaning/. Ch16 n.5
- Jeff Speaks. "Transparency, Intentionalism, and the Nature of Perceptual Content." *Philosophy and Phenomenological Research* 79, no. 3 (2009): 539–573. Transparency n.6
- Mark Sprevak. "Three Challenges to Chalmers on Computational Implementation." *Journal of Cognitive Science* 13 (2012): 107–143. Ch22 n.5
- Alison Ann Springle. "Perception, Representation, Realism, and Function." *Philosophy of Science* 86, no. 5 (2019): 1202–1213. DOI 10.1086/705525. Ch14 n.3
- Daniel Stoljar. "Physicalism and Phenomenal Concepts." *Mind & Language* 20, no. 5 (2005): 469–494. Ch11 n.8
- Daniel Stoljar. "The Argument from Diaphanousness." *Canadian Journal of Philosophy*, supplementary vol. 30, *New Essays in the Philosophy of Language and Mind* (2004): 341–390. doi:10.1080/00455091.2004.10717610. Ch04 n.4, Ch06 n.3, Ch06 n.9, Transparency n.4
- Daniel Stoljar. "Two Conceptions of the Physical." *Philosophy and Phenomenological Research* 62, no. 2 (2001): 253–281. Ch10 n.1
- Galen Strawson. "Against Narrativity." In *Real Materialism and Other Essays* (Oxford: Clarendon Press, 2008), 189–208. Ch17 n.13
- Galen Strawson. "Real Direct Realism: Reflections on Perception." In *Phenomenal Qualities: Sense, Perception, and Consciousness*, ed. Paul Coates and Sam Coleman (Oxford: Oxford University Press, 2015), 214–253. DOI 10.1093/acprof:oso/9780198712718.003.0008. Ch02 n.8

- Galen Strawson. “Real Materialism.” In *Real Materialism and Other Essays* (Oxford: Clarendon Press, 2008), 19–52. DOI 10.1093/acprof:oso/9780199267422.003.0002. Ch06 n.6
- Galen Strawson. “Realistic Monism: Why Physicalism Entails Panpsychism.” *Journal of Consciousness Studies* 13, nos. 10–11 (2006): 3–31. Ch10 n.3
- Galen Strawson. “The Impossibility of Moral Responsibility.” *Philosophical Studies* 75, nos. 1–2 (1994): 5–24. DOI 10.1007/BF00989879. Ch18 n.3
- P. F. Strawson. “Freedom and Resentment.” *Proceedings of the British Academy* 48 (1962): 1–25. Ch17 n.12, Ch18 n.4
- P. F. Strawson. *The Bounds of Sense: An Essay on Kant’s “Critique of Pure Reason”* (London: Methuen, 1966). Ch12 n.13
- Anders Søgaard. “Understanding Models Understanding Language.” *Synthese* 200 (2022): article 443. DOI 10.1007/s11229-022-03931-4. Ch23 n.4, Ch24 n.11, Ch24 n.14
- Qwen Team. “Qwen3 Technical Report.” arXiv:2505.09388v1 (2025). Ch20 n.3
- Adly Templeton et al. “Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet.” *Transformer Circuits Thread*, May 21, 2024. Ch23 n.10b
- Corbett H. Thigpen and Hervey M. Cleckley. *The Three Faces of Eve* (New York: McGraw-Hill, 1957). Ch17 n.10
- Evan Thompson. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind* (Cambridge, MA: Belknap Press of Harvard University Press, 2007). App. B n.39
- Nitasha Tiku. “The Google engineer who thinks the company’s AI has come to life.” *The Washington Post* June 11 (2022). Preface n.1
- Giulio Tononi and Christof Koch. “Consciousness: Here, There and Everywhere?” *Philosophical Transactions of the Royal Society B* 370 (2015): 20140167. App. B n.61
- Giulio Tononi, Melanie Boly, Marcello Massimini, and Christof Koch. “Integrated Information Theory: From Consciousness to Its Physical Substrate.” *Nature Reviews Neuroscience* 17, no. 7 (2016): 450–461. doi:10.1038/nrn.2016.44. Ch06 n.21d, App. B n.60
- Charles Travis. “The Silence of the Senses.” *Mind* 113, no. 449 (2004): 57–94. Ch06 n.17, App. B n.74
- Judy Trevena and Jeff Miller. “Brain Preparation Before a Voluntary Action: Evidence Against Unconscious Movement Initiation.” *Consciousness and Cognition* 19, no. 1 (2010): 447–456. DOI 10.1016/j.concog.2009.08.006. App. B n.63
- Endel Tulving. *Elements of Episodic Memory* (Oxford: Clarendon Press, 1983). Ch02 n.9
- Alan M. Turing. “Computing Machinery and Intelligence.” *Mind* 59, no. 236 (1950): 433–460. DOI 10.1093/mind/LIX.236.433.

- Sherry Turkle. *Alone Together: Why We Expect More from Technology and Less from Each Other* (New York: Basic Books, 2011). Ch25 n.2
- Sherry Turkle. *Life on the Screen: Identity in the Age of the Internet* (New York: Simon & Schuster, 1995). Ch25 n.2
- Kasimir Twardowski. *On the Content and Object of Presentations: A Psychological Investigation*, trans. and intro. R. Grossmann (The Hague: Martinus Nijhoff, 1977). Ch01 n.1
- Michael Tye. "Another Look at Representationalism about Pain." In *Pain: New Essays on Its Nature and the Methodology of Its Study*, ed. Murat Aydede, 99–120 (Cambridge, MA: MIT Press, 2005). DOI 10.7551/mitpress/5211.003.0005. Ch07 n.2
- Michael Tye. "How Can We Tell If a Machine Is Conscious?" *Inquiry* 69, no. 6 (2026): 3040–3058. DOI 10.1080/0020174X.2024.2434856. Ch24 n.15
- Michael Tye. "How I Learned to Stop Worrying and Love Panpsychism." *Journal of Consciousness Studies* 31, nos. 9–10 (2024): 10–28. doi:10.53765/20512201.31.9.010. Ch10 n.5, Ch10 n.6, Ch10 n.10, Inverted Earth n.20, Ch10 n.7, App. B n.40
- Michael Tye. "Intentionalism and the Argument from No Common Content." *Philosophical Perspectives* 21, no. 1 (2007): 589–613. doi:10.1111/j.1520-8583.2007.00137.x. App. B n.14
- Michael Tye. "Nonconceptual Content, Richness, and Fineness of Grain." In *Perceptual Experience*, ed. Tamar Szabó Gendler and John Hawthorne (Oxford: Oxford University Press, 2006), 504–530. Ch06 n.20
- Michael Tye. "Phenomenal Consciousness: The Explanatory Gap as a Cognitive Illusion." *Mind* 108, no. 432 (1999): 705–725. DOI 10.1093/mind/108.432.705. Ch11 n.8
- Michael Tye. "Qualia." *Stanford Encyclopedia of Philosophy*, substantive revision September 19, 2025. App. B n.41
- Michael Tye. "Representationalism and the Transparency of Experience." *Noûs* 36, no. 1 (2002): 137–151. DOI 10.1111/1468-0068.00365. Transparency n.8
- Michael Tye. "The Experience of Emotion: An Intentionalist Theory." *Revue internationale de philosophie*, no. 243 (2008/1): 25–50. DOI 10.3917/rip.243.0025. Ch07 epigraph
- Michael Tye. "The Truth about True Blue." *Analysis* 66, no. 4 (2006): 340–344. DOI 10.1093/analys/66.4.340. App. B n.22
- Michael Tye. "Transparency, Qualia Realism and Representationalism." *Philosophical Studies* 170, no. 1 (2014): 39–57. DOI 10.1007/s11098-013-0177-8. Ch06 nn.7, 9
- Michael Tye. "Yes, Phenomenal Character Really Is Out There in the World." *Philosophy and Phenomenological Research* 91, no. 2 (2015): 483–488. DOI 10.1111/phpr.12167. Ch06 n.20
- Michael Tye. *Consciousness and Persons: Unity and Identity* (Cambridge, MA: MIT Press, 2003). Ch06 n.21b

- Michael Tye. *Consciousness Revisited: Materialism Without Phenomenal Concepts* (Cambridge, MA: MIT Press, 2009). Ch06 n.18a, Ch08 nn.11, 14, Ch11 n.8
- Michael Tye. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind* (Cambridge, MA: MIT Press, 1995). Ch02 n.9, Ch04 n.6, Ch06 nn.4, 14a, 20, Ch07 nn.2, 10, Ch09 n.4, Ch10 n.5, Ch20 n.18, Ch22 n.12, Ch24 n.15, App. B nn.14, 25, 27, 34
- Michael Tye. *Vagueness and the Evolution of Consciousness: Through the Looking Glass* (Oxford: Oxford University Press, 2021). DOI 10.1093/oso/9780198867234.001.0001. Ch10 nn.5, 7, App. B n.34, Inverted Earth n.20
- Peter van Inwagen. *An Essay on Free Will* (Oxford: Clarendon Press, 1983). App. B n.65
- Francisco J. Varela, Evan Thompson, and Eleanor Rosch. *The Embodied Mind: Cognitive Science and Human Experience* (Cambridge, MA: MIT Press, 1991). Ch15 n.11, Ch21 n.4
- Manuel Vargas. *Building Better Beings: A Theory of Moral Responsibility* (Oxford University Press, 2013). DOI 10.1093/acprof:oso/9780199697540.001.0001. Ch18 n.5
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need." In *Advances in Neural Information Processing Systems* 30 (2017). https://papers.nips.cc/paper/7181-attention-is-all-you-need. Ch23 n.6
- J. David Velleman. "The Self as Narrator." *Self to Self: Selected Essays* (Cambridge: Cambridge University Press, 2006). Ch17 n.10
- Johannes von Oswald et al. "Transformers Learn In-Context by Gradient Descent." arXiv:2212.07677 (2022). Ch23 n.10b
- L. S. Vygotsky. *Mind in Society: The Development of Higher Psychological Processes*, ed. Michael Cole, Vera John-Steiner, Sylvia Scribner, and Ellen Souberman (Cambridge, MA: Harvard University Press, 1978). Ch17 n.2
- Lev Vygotsky. *Thought and Language*, trans. Eugenia Hanfmann and Gertrude Vakar (Cambridge, MA: MIT Press, 1962 [1934]); revised edition, trans. and ed. Alex Kozulin (Cambridge, MA: MIT Press, 1986). Ch17 n.2, App. B n.1
- R. Jay Wallace. *Responsibility and the Moral Sentiments* (Cambridge, MA: Harvard University Press, 1994). Ch18 n.5
- Zina B. Ward. "Directive Representation and the Job Description Challenge." *Philosophy and the Mind Sciences* 7 (2026): 1–19. DOI 10.33735/phimisci.2026.12005. Ch15 n.5
- Gary Watson. "Free Agency." *The Journal of Philosophy* 72, no. 8 (1975): 205–220. DOI 10.2307/2024703. Ch18 n.5
- Adam Waytz, John Cacioppo, and Nicholas Epley. "Who Sees Human? The Stability and Importance of Individual Differences in Anthropomorphism." *Perspectives on Psychological Science* 5, no. 3 (2010): 219–232. Ch25 n.4

- Jason Wei et al. “Emergent Abilities of Large Language Models.” *Transactions on Machine Learning Research* (August 2022). https://openreview.net/forum?id=yzkSU5zdwD. Ch21 n.8
- Daniel Weintraub et al. “Impulse Control Disorders in Parkinson Disease: A Cross-Sectional Study of 3090 Patients.” *Archives of Neurology* 67, no. 5 (2010): 589–595. doi:10.1001/archneurol.2010.65. Ch18 n.1
- Joseph Weizenbaum. “ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine.” *Communications of the ACM* 9, no. 1 (1966): 36–45. DOI 10.1145/365153.365168. Ch19 n.5
- Joseph Weizenbaum. *Computer Power and Human Reason: From Judgment to Calculation* (San Francisco: W. H. Freeman, 1976). Ch19 n.5, Ch25 n.2
- Michael Wheeler. *Reconstructing the Cognitive World: The Next Step* (Cambridge, MA: MIT Press, 2005). Ch20 n.24
- Wanja Wiese. “Artificial Consciousness: A Perspective from the Free Energy Principle.” *Philosophical Studies* 181, no. 8 (2024): 1947–1970. DOI 10.1007/s11098-024-02182-y. Ch24 n.10
- Bernard Williams. “The Makropulos Case: Reflections on the Tedium of Immortality.” In *Problems of the Self* (Cambridge: Cambridge University Press, 1973), 82–100. Ch24 n.9
- Iwan Williams. “Can Structural Correspondences Ground Real-World Representational Content in Large Language Models?” *Mind & Language*, published online March 7, 2026. DOI 10.1111/mila.70018. Ch23 n.8
- J. Robert G. Williams. “Eligibility and Inscrutability.” *The Philosophical Review* 116, no. 3 (2007): 361–399. DOI 10.1215/00318108-2007-002. Ch22 n.20
- Ludwig Wittgenstein. *Philosophical Investigations*, trans. G. E. M. Anscombe, 3rd ed. (Oxford: Basil Blackwell, 1967; English-text reprint, 1972). Ch11 n.11, Ch16 n.7, Ch16 n.8, Ch17 n.5, App. B n.1, Ch19 (epigraph)
- Ludwig Wittgenstein. *Remarks on the Philosophy of Psychology*. Vol. 2, ed. G. H. von Wright and Heikki Nyman, trans. C. G. Luckhardt and M. A. E. Aue (Oxford: Basil Blackwell, 1980). Ch19 n.4
- Ludwig Wittgenstein. *Zettel*, ed. G. E. M. Anscombe and G. H. von Wright, trans. G. E. M. Anscombe, 2nd ed. (Oxford: Basil Blackwell, 1981). Ch19 n.4
- Susan Wolf. *Freedom within Reason* (Oxford: Oxford University Press, 1990). Ch18 n.5
- Stephen Yablo. “Coulda, Woulda, Shoulda.” In *Conceivability and Possibility*, ed. Tamar Szabó Gendler and John Hawthorne (Oxford: Oxford University Press, 2002), 441–492. DOI 10.1093/oso/9780198250890.003.0014. Ch11 n.12
- Helen Yetter-Chappell. “Dissolving Type-B Physicalism.” *Philosophical Perspectives* 31, no. 1 (2017): 469–498. doi:10.1111/phpe.12099. Ch11 n.18
- Helen Yetter-Chappell. “Guessing at Ghosts in the Machine.” *Ratio* 39, no. 2 (2026): 73–81. DOI 10.1111/rati.70013. Ch25 n.11

- John W. Yolton. *Perception & Reality: A History from Descartes to Kant* (Ithaca: Cornell University Press, 1996). Ch02 n.4
- John W. Yolton. *Perceptual Acquaintance: From Descartes to Reid* (Minneapolis: University of Minnesota Press, 1984). Ch02 n.4
- Dan Zahavi. "Beyond Empathy: Phenomenological Approaches to Intersubjectivity." *Journal of Consciousness Studies* 8, nos. 5–7 (2001): 151–167.
- Dan Zahavi. "Empathy and Direct Social Perception: A Phenomenological Proposal." *Review of Philosophy and Psychology* 2, no. 3 (2011): 541–558. DOI 10.1007/s13164-011-0070-3. Ch19 n.3
- Jenny Zhang et al. "Darwin Gödel Machine: Open-Ended Evolution of Self-Improving Agents." arXiv:2505.22954 (2025). App. B n.51b