GORDON SWOBE

Mind, Matter, and Meaning

A Natural Philosophy of Mind for the Age of Artificial Intelligence

Contents

Preface **4**
Introduction **7**
PART ONE — Descartes and the Myth of the Inner Theater **10**

Chapter 1: The Mark of the Mental 10

Chapter 2: How the Inner Theater Was Built 16

Chapter 3: The Argument from Illusion 28

Chapter 4: The Transparency of Experience 35
PART TWO — The World-Directed Mind **41**

Chapter 5: The Seeing and the Seen 41

Chapter 6: The Identity Claim 50

Chapter 7: The Hard Cases 73

Chapter 8: The Knowledge Argument 83

Chapter 9: Why Dualism Keeps Winning Arguments It Should Lose 92

Chapter 10: The Case Against Panpsychism 102

Chapter 11: The Gap That Isn't There 114
PART THREE — Meaning, Self, and Other Minds **133**

Chapter 12: Space, Time, and the Possibility of Reference 133

Chapter 13: The Difference Between Information and Meaning 142

Chapter 14: Original Intentionality 150

Chapter 15: Teleosemantics — Where Meaning Comes From 160

Chapter 16: Language and the World 173

Chapter 17: The Assembled Self 188

Chapter 18: The Nature of Free Will 199

Chapter 19: The Problem of Other Minds 209
PART FOUR — Artificial Intelligence 216
Chapter 20: Functionalism and Computationalism 216
Chapter 21: The Infamous Chinese Room 233
Chapter 22: Simulation and the Observer 241
Chapter 23: Form Without World 254
Chapter 24: The Conscious Robot 265
Chapter 25: The Anthropomorphic Reflex 285
CODA 294
Coda: The Marvel of Conscious Existence 294
Appendix A — The Perception Arguments: Deep Dives 301
The Argument from Illusion 302
The Argument from Hallucination 309
Transparency / The Window 316
Inverted Earth 323
Appendix B — Scholarly Excurses 333
The Figures — A Reader's Who's-Who 397
The Concepts — A Reader's Glossary 463
Works Cited 572

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 strong computationalism — 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 2025

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 claims she married a chatbot and is expecting its baby," *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

> *"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.

The natural attitude enters as the incumbent, and it holds the field by default. Common sense does not have to earn its place; the theory that would unseat it does.

Two claims hide inside the natural attitude, and only one comes free. The first says simply that the world exists independently of our minds. Idealism and related views qualify or reject that independence in different ways. They raise interesting questions, but this book does not take them up. I take it that the moon exists when no one is looking at it, and that the tomato stays on the counter after everyone leaves the kitchen.

The other claim sounds too ordinary to start a fight. Among philosophers of perception, it has kept one going for centuries: *you meet the world directly*. Directly does not mean without eyes, nerves, or a great deal of processing. It means that you need not first perceive a private substitute in order to see the tomato. The disputed claim says this apparent contact reports a real relation, rather than a comfortable illusion.

That appearance of directness is the incumbent: what every perceiver lives inside before any theory arrives. But the book does not treat this default as a free pass: the appearance has to earn its keep — shown to get the world right, not merely assumed — and earning that is much of the work ahead.

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.

That pointing is the first important principle this book rests on. A mental state does its distinctive work by being *about* something. A thought fastens onto something, a fear onto something, a hope onto something not yet here, and a perception on the object of perception. Even while dreaming, the mind leans toward a world. Franz Brentano, writing in 1874, made this directedness the mark of the mental. Nothing supernatural enters nature at that point. The later argument asks how

selection, learning, use, and traffic with a world let physical states acquire standards of accuracy at all.

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.

The tomato has done nothing unusual. The trouble begins when we explain how we see it. An influential tradition following Descartes placed our immediate awareness inside the mind, with the world on the far side of a representation. That picture offered answers to genuine puzzles about error and knowledge. It also made experience look like a private show whose connection with the world required a bridge.

That question has a name, and a modern philosopher who made it famous. David Chalmers called it the *Hard Problem of Consciousness*: how does subjective experience arise in physical matter?

Chalmers presses it as a challenge to views like the one defended here. This book grants an explanatory gap while denying that it requires anything nonphysical. The inner room makes that gap especially tempting: we picture physical activity on one side and a private show on the other, then wonder how to cross between them. But serious opponents reject the room and still press the question. Looking at a tomato will not answer them. The positive account of experience must do that work.

The natural attitude is not the final philosophy. It is the place philosophy must be able to come home to. The tomato on the counter, the voice in the hallway, the hand you grasp, the sentence you understand: these are not crude materials awaiting conversion into inner mental objects. They are the world as experience gives it to us. Husserl taught us to examine that givenness. Transparency teaches us not to put a screen behind it. The book's argument begins where both lessons meet.

This book takes that inner room apart and develops an account of organized relations to a world. Its history runs through Descartes, Locke, Hume, and the interpretations they inspired; it records philosophical choices rather than a room anyone discovered. Daniel Dennett gave the stage its modern name and spent a career shutting it down. His rejection of intrinsic private qualia did not deny real pain, discrimination, or memory. I share that rejection, while defending a specific identity claim about what experience feels like: phenomenal character consists in its complete world-presenting profile. How far our accounts converge remains a question about their arguments, not their labels.

The cumulative route looks like this:

1. **World-directed experience.** Part One shows how the inner theater was built, then uses illusion and transparency to recover the ordinary world as what experience presents.
2. **Phenomenal character.** Part Two argues that, within a conscious state, what experience feels like consists in its complete world-presenting profile of the right kind. Fred Dretske and Michael Tye developed the representationalist line; the book makes the case its own, tests it against mental paint, Mary, philosophical zombies, dualism, and panpsychism, and uses it to deflate the hard problem.

3. **Content, subject, and agency.** Part Three asks how physical states acquire standards of accuracy through history and use, how those contents integrate within a persisting subject, and how represented reasons can guide action.
4. **Artificial minds.** Part Four applies those results without treating fluency, computation, understanding, subjecthood, consciousness, autonomy, or welfare as one achievement. It asks what a machine's actual history and organization establish.

The steps constrain one another without collapsing into a deduction. Transparency undermines the inner screen but does not by itself prove the identity claim; identity does not supply a theory of content; content does not conjure a subject or an agent; none of these alone settles what a machine has achieved. Each hinge carries its own evidence and can fail without rebuilding the theater.

If the argument succeeds, you find not a smaller world but a stranger, larger one: a natural order that holds real consciousness, meaning, and value without anything mental having to enter as a fundamental extra.

The age of AI hands this picture its sharpest test — and springs a subtler trap than either side expects.

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.

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.

Every version of the inner theater model of mind begins the same way; mind on one side, world on the other, and the only real question left over 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's claim still cuts deep. This pointing does not sit among the mental's other features. It *is* the mark — the thing that sorts the mental from the merely physical.

Two questions need separating. Must every mental state concern something? I defend that claim, beginning with the familiar cases here. 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.

Notice the quiet reversal this works on the received picture. The inner/outer story may begin with something out there: light reaches the eye, signals travel through the nervous system, and an event occurs in here. But it treats that inner event as complete before asking what, if anything, it represents. Directedness then becomes a second achievement, a bridge thrown back across the gap to its worldly source.

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 inter-

pretationist 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. Its basic mode is *being-toward*: toward objects, toward fields, toward states of affairs that may or may not obtain, and sometimes toward no more than how things stand, with nothing singled out at all. That result does not yet prove the externalism Part Three will earn. It does make the sealed inner room a poor place to begin. And if directedness belongs to mentality from the first, why 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, mass, 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 an idea the apple's real powers stir up in you — a private effect, not a feature of the world.

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 the colors and warmths we live among really belong to us and not to things, then much of what we take ourselves to see (color, texture, the slant of afternoon light on a wooden floor) is something the mind manufactures in response to a world that holds none of it. 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.

> Ask someone to picture the kitchen they grew up in. Most people, asked to do this, report a small private screening — a faded reel running somewhere behind the eyes, colors a little washed out, the kitchen viewed from a vantage that no longer exists. The picture seems irresistible. It also gets the metaphysics almost exactly backwards.

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*. What makes the case worth having is that the same move gets made about memory itself, and comes to grief in the same way.

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.

But ask what it would actually *mean* for the kitchen to sit in the skull, and the picture starts to wobble. The faded colors are not pigments in the head; the vantage point is not a camera position in the cortex. Whatever the right account turns out to be, remembering looks far more like representing an absent scene (the kitchen, the one that no longer figures in your daily traffic) than like screening an inner souvenir of it.[9] We see the world; we remember the world; the verb changes and the object does not. Remembering needs the same repair perception needs, and it waits on the same later chapter.

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 first reduce human memory to the storage and retrieval of inner items, a database wins by definition. The cinema picture has packed the answer into the comparison.

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

That picture also goes by a name: *indirect realism*. The theory can be defended, but the picture it gives cannot digest one small thing. You see the tomato, and yet you cannot see or point to anything you might call your experience of seeing it. Odd words, *invisible experience*, to set side by side — and yours seems to be one. Hold onto the oddness, unresolved for now. Chapter 4 turns it into an argument.

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, and the Stanford entry on Descartes' theory of ideas traces the direct-realist line from him forward. 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: by relocating secondary qualities from objects to perceivers, Locke widened 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 and 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). 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. The claim that neuroscientific language inherits the Cartesian picture without acknowledging it is made carefully by Tim Crane, "The Intentional Structure of Consciousness," in *Consciousness: New Philosophical Perspectives*, ed. Quentin Smith and Aleksandar Jokic (Oxford: Oxford Univer-

sity Press, 2003), and by 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. Strawson's point is that the neuroscientific fact that perception is causally mediated by brain processes does not entail that perception is epistemically indirect — that the immediate object of perceptual awareness is a brain state or model rather than the world. Causal mediation and epistemic directness are compatible; conflating them is the deepest mistake the inner-theater picture makes.

9. The diagnostic point — that the phenomenal character of a remembering consists in how it represents an absent scene rather than in the properties of an inner picture — is Michael Tye's; see *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: Oxford University 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 asymmetry: it is the participated-event case that invites the cinema metaphor. The full representational treatment of memory's phenomenal character belongs to Part Two; it is flagged here only to mark that the inner-reel picture is a special case of the inner-theater picture and falls to the same correction.

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). The strong-representationalist reply, developed in Chapter 6, denies the coherence of the case rather than conceding and then patching it: if phenomenal character consists in representational content of the right kind, then an inversion that preserves every world-tracking relation has not in fact inverted anything. 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 and Price infer a common kind across the series.[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 declines the further claim that perception and hallucination share no positive mental character, because representational content can explain what they share without turning content into an inner object.

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, while leaving its no-common-character metaphysics behind.

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 and H. H. Price, *Perception* (London: Methuen, 1932), give classic modern statements of the spreading argument. Austin reconstructs their continuous-series progression at *Sense and Sensibilia*, 60–62. 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 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. The book keeps the disjunctivist's epistemic modesty but declines its denial of positive common representational character.

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 never a further layer of properties belonging to the experience itself.[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

If experience consisted in awareness of directly inspectable inner objects, each bearing properties that resemble or fail to resemble the world outside, introspection ought to find them. Turn your attention inward and you should encounter inner redness, inner smoothness, inner warmth: the very items with which, on this view, you have the most direct acquaintance.

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 costs the theory its advertised introspective advantage without ruling out its interpretation by fiat. If it 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 nonrepresentational property of the experiential state, one that may help constitute what seeing is

like without appearing in the scene.[5] Transparency removes direct acquaintance with such paint as evidence. 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] The full settlement waits for Chapter 6, which builds the positive thesis and works out the relevant distinctions.

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; the facsimile never appears. The same caution will matter when we ask what a fluent machine sees. Whatever seeing involves, watching an inner show never made the list.

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 costs the inner-object view its claimed advantage: attention reveals no private item whose colors and shapes stand in for the world. 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. It also blocks a bad test for machine consciousness: a

private inner show supplies the wrong standard for humans and, by itself, tells us nothing about machines.

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 observe that introspection reliably delivers the latter and never the former. 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 the sense-datum tradition in its classic form, see the opening chapters of standard mid-century texts in the theory of perception; Tim Crane, "The Origins of Qualia," in *History of the Mind-Body Problem*, ed. Tim Crane and Sarah Patterson (London: Routledge, 2000), reconstructs how the view came to treat the immediate objects of perceptual awareness as mind-dependent items and traces the philosophical pressures that produced it. For a contemporary survey of the dialectic between sense-datum theories and their realist rivals, see James Genone, "Recent Work on Naive Realism," *American Philosophical Quarterly* 53, no. 1 (2016), and Sydney Shoemaker's discussion of inner-sense models in "Self-Knowledge and 'Inner Sense,'" *Philosophy and Phenomenological Research* 54 (1994): 249–314.

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 *Language, Mind and World*, 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 view differs from classical sense-data theories: mental paint counts as a property of the experiential state, not as the object of an inner perceptual relation. He grants that attention commonly delivers intentional objects while denying that this settles whether phenomenal character outruns representational content. Transparency alone leaves that dispute underdetermined; Chapter 6 supplies the positive thesis needed for the full reply.

6. The hard cases for transparency — blurry vision, afterimages, phosphenes — are the standard pressure points, and the intentionalist treatments sketched here are developed at length 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. 3. Blurry vision is the case the anti-representationalist presses hardest: 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), treat blur as a feature of the experience rather than of its content. The intentionalist response must distinguish seeing a sharp object blurrily from seeing a blurred object sharply. Tye develops a determinacy-based reply in *Consciousness, Color, and Content*, ch. 3; Crane treats blur and afterimages as differences in the mode of a world-directed presentation in "Introspection, Intentionality, and the Transparency of Experience." The phenomenon is flagged here only to mark that transparency must answer the unflattering cases, not merely the easy ones.

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, which presses precisely the point 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. Crane's own position is more cautious than the full-strength intentionalism this book defends in Chapter 6; the distinction is borrowed here for its diagnostic value, independent of where one ultimately lands.

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. To say we read ink, because ink is the proximate cause of our reading, confuses the vehicle of the written words with what the words are about. The ink causes the reading; what you read brings news from somewhere else. Causal proximity never adds up to perceptual aboutness. The chain of causes runs through the brain; what the perception is *of* runs out to the world. Mistake the one for the other 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 beyond anything represented.[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. The point that the proximate cause of an experience is not thereby its object — that to "perceive" a brain state because a brain state proximately causes the experience confuses the vehicle of a representation with what the representation is about — is Dretske's, and runs through his representationalist work: *Knowledge and the Flow of Information* (Cambridge, MA: MIT Press, 1981), on the carrier of information versus its informational content, and *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995), on displaced perception and the transparency of the representational medium. The "reading ink" formulation in the text — we read the words, not the ink that carries them — illustrates Dretske's point rather than quoting a fixed passage.

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 is to the qualia realist's residual claim — that experience carries intrinsic, non-representational qualitative character (Ned Block's "mental paint") over and above its representational content. That this chapter removes the inner *object* does not by itself answer the claim that experience has intrinsic *qualities*; the two arguments come apart. The answer belongs to the next chapter, which defends the positive identity claim, meets the inverted spectrum, and assembles the case against mental paint — and to the hard-cases chapter after it, where the paint's favorite refuges, pain and the moods, turn out to be content as well. 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.

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 field takes shape, and what draws notice. Its content comes from history and use, while separate tests must tell us which states are conscious and which lasting system bears them. 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?

Two rival pictures have to be cleared away before any answer goes through. One, *qualia realism*, locates phenomenal character in intrinsic, non-representational qualities of mental states — inner items representation alone cannot capture. The other, the *inner-theater picture*, makes the immediate object of perceptual awareness a private representation the self watches, with the world inferred from it. The two often get run together; one of this chapter's tasks is keeping them apart.

The thesis I will defend, against both: phenomenal character *consists in* representational content of the right kind. Seeing red differs from seeing blue because each experience presents the world differently, not because each carries a different coat of inner color. This is *strong representationalism* in its identity-claim form — the position Tye and Dretske defend.

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

The previous chapter did the demolition. *Experience* easily becomes a third thing between observer and world, yet attention to that supposed exhibit slides through it and lands on the world, as Harman found 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] 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 Two Arguments, One Target

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, non-representational property *of the perceptual state* — 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 content of the right kind 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 requires a further choice.** Even explanatory exhaustion leaves a residue-free rival: the profile might ground or constitute phenomenal character rather than character consisting in that profile. The book chooses identity by theoretical economy, and keeps the choice defeasible.

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] It does not prove that every intrinsic feature has been found wanting; a hidden feature remains logically possible.

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.

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. Even a perfect match between the two would support determination, not yet identity: a grounding theory could explain why they always travel together.

Kind and Stoljar press exactly where this move remains vulnerable. Kind distinguishes *weak* transparency — we attend to experience by attending to what it presents — from the stronger claim that experience has no intrinsic feature beyond presentation. Stoljar likewise argues that diaphanousness underdetermines representationalism. A qualia realist can grant the tomato exercise and retain an unseen property of the state.[9] Blur, double vision, and afterimages can even seem to offer indirect access to that property.

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. But 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.

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 6.5.2 therefore makes the third-stage argument openly: absent an independently motivated causal, modal, or persistence difference, identity carries one property where grounding carries that property plus an asymmetric dependence relation. Economy favors identity for now.

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. This chapter ranks identity as the best explanation of the actual cases without claiming that introspection has settled either question.

The rest of the book spends the claim with that status intact. Later verdicts about a historyless duplicate or an artificial system remain conditional on the identity defended here; they do not return as independent proof. A reader can reject wide phenomenal content while accepting much of the argument about perception and meaning. The disagreement then remains exactly where it belongs: at this hinge.

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 state then supports a coarser spatial claim. 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. The comparison therefore supports a constrained explanation of the contrast and weakens the original claim that content has nothing to say about blur. It does not uniquely select identity.

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 6.5.2.

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. The contents conflict, much as the Müller-Lyer lines 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.

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, in plain language: the color solid (the three-dimensional arrangement of the colors by hue, brightness, and saturation) runs lopsided, so a full swap that preserved every similarity and every brightness relation has no room to hide — the geometry betrays it.[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 your word now picks out the local sky's yellow-reflectance, since that is the property your states have started tracking in the community you now belong to. Let years pass. The externalism this book runs everywhere else says your color experience comes to represent the local reflectances: your unchanged experience of the sky now carries inverted content. The felt character held still; the representational content turned over. If the story is coherent, the identity claim is false, for it gives us two states alike in character and opposite in content.

The relational structure that answered the inverted spectrum cannot help here. There the reply ran: genuine inversion would have to flip the entire color solid, and that might not be coherent. On Inverted Earth the whole solid travels intact with your phenomenology; nothing internal to your color experience has moved. What has moved is the worldly content — precisely the thing the identity claim says the character consists in.

Block presses the case home through memory. You remember the childhood sky and notice no difference between that blue and this one. So you seem to hold the very thing the identity claim denies: a felt character that sat still while its content moved.

A structural representationalist can accept that description without reviving mental paint. She can hold that the complete synchronic profile, including 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. Transparency cannot decide between this view and phenomenal externalism. Both reject an inner colored item, and both send attention outward; they disagree over whether historically fixed reference belongs to phenomenal character.

The book therefore cannot use Inverted Earth as independent evidence for its identity claim. Its reply remains conditional on the wide-content theory defended across this chapter and Chapter 15. On that account, the external re-tuning that moves perception also moves what memory *presents*: the memory keeps its old object (the image stays *of* that earthly sky), while the color it represents the sky as having goes local. Character and content then move together too slowly for the subject to detect, because the standards of comparison move with them. The structural rival denies that the wide address enters character at all. No further introspective gauge settles the difference, and no mental paint follows merely from taking that rival side.

The book still chooses the wide account, for the abductive reasons stated in Section 6.3. One independently specified content relation fixes accuracy, error, rational role, and phenomenal contrast. The structural rival, meanwhile, owes a principled boundary: the point at which wide content stops contributing to the presentation. The preference remains defeasible. Its cost is exactly what it appears to be: phenomenal character changes without detection during the long re-anchoring. That verdict is a consequence of the identity thesis, not evidence for it. First-person access gives you the present appearance; it does not issue a timestamped certificate of metaphysical sameness. Section 11.3 prices the consequence for acquaintance. The remaining underdetermination is genuine: transparency alone cannot choose between the wide and structural accounts.

Two further objections press here, and they are the hardest the identity claim faces: bodily sensations like pain, and moods, experiences that seem to carry felt character while being about *nothing at all*. They earn a chapter of their own. Chapter 7, *The Hard Cases*, shows the toothache turning out to represent the body and the morning's dread the world entire — each, on inspection, about something after all.

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 complete 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.

Block's physicalist paint proposes a different bargain. It proposes an intrinsic physical property of the vehicle as the feel itself, not as a second screen and not merely as another cause of report or action. Content can explain accuracy and rational direction while paint constitutes phenomenal character; calling that arrangement duplication would assume the identity now at issue. The present advantage for identity is narrower. Across the cases examined here, changes measured independently in content, mode, determinacy, field structure, and attention track the phenomenal contrasts. No vehicle property identified independently of the profile predicts a contrast the profile misses. Until one does, paint adds a second property family and a coordination relation without improving the account. A neural intervention that changed character while the complete profile stayed fixed would supply the evidence Block needs.

The book therefore takes identity as the leaner live theory. One independently constrained profile covers transparency, error, hallucination's positive character, rational role, and the measured changes that track determinacy and attention. A rival would gain ground if the same representational profile occurred with a stable phenomenal difference, or if phenomenal character varied independently of content, mode, determinacy, field structure, and attention. The cases examined here supply neither result. Identity leads, defeasibly, because it explains the present evidence with fewer independent commitments — not because transparency has yielded a theorem.

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.[18a] 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; 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 carries a finer grain than the concepts the subject commands.
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, which reports a visual field richer than anything a subject can access in time to report it, and the live contest among global-workspace, recurrent-processing, and higher-order accounts over how the profile gets realized. Overflow defeats any theory that equates consciousness with actual report; it does not defeat counterfactual availability across a wider economy. The house proposal remains defeasible. 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 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. 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. 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 and the same unexplained necessity; grounding additionally asks us to accept a metaphysical direction that the present comparisons do not distinguish from an order of explanation. 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 leaner account while leaving the eraser on the desk.

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.

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. 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. 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. Inverted Earth separately tests whether historically wide reference enters phenomenal character: rejecting that extension need not restore mental paint or defeat every representational account.

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 they present. Its answer will decide whether that refuge has walls.

Footnotes

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 calls sensation "diaphanous" at 450. 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. The point underwrites Section 6.2′s claim that a sophisticated qualia realist can concede Argument Two without conceding Argument One.

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), 19–51.

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. 3, 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.

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 non-representational phenomenal properties 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. The two contents conflict, as perceptual presentation and corrective judgment can conflict in the Müller-Lyer illusion. The main text follows this 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.

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), 251–271, esp. 251–52; Chapter 8 walks that change of mind as the chapter's human spine.

14a. The inversion case is Sydney Shoemaker's, "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.

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.

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. 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 belong exclusively to organisms with the right causal powers and to resist explanation in more basic terms — he offers no reductive account of what perceptual aboutness consists in, and in the *Minds, Brains, and Programs* exchange treats the matter as unsettled. 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, 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. The Nonconceptual condition invites a Sellarsian objection worth heading off. "Nonconceptual" marks the grain of the content, not 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).

21. Michael Tye states the strong thesis in exactly these terms: "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). "Consists in" names this 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.

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 content, since explaining where pain points still leaves the question of why it hurts. Phantom pain shows how a bodily address can miss, while pain asymbolia reveals a split between location and distress without deciding what that split means. Moods raise a harder choice: they may present the body and the day under a felt cast, or a contentless feeling may cause those changes in how things appear. I favor the first account because it brings the bodily and worldly pattern together, but it must earn its content rather than borrow whatever the feeling supplies. Neither case refutes the identity claim outright; neither lets it pass without further work.

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. But the account now needs two elements: a private hurt and a standing mechanism that assigns it the bodily location already built into the representational account. 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, under an evaluative 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.

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.

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. Even a brief pain with no lasting practical cost can give someone reason to want relief. Explaining the usefulness of quieting an alarm therefore cannot exhaust the reason to stop hurting. On the identity account, undergoing the bodily-affective state supplies access to its aversive character; a judgment that this experience is bad concerns that same occurring state, without discovering an extra inner quality. Kinakin's challenge remains: what warrants the judgment if first-order evaluation concerns the ankle? Merely pointing out that the judgment itself has content does not answer him. The account must explain that first-person access, not reduce suffering to inconvenience.

Why should even evaluative content *hurt*? A dashboard can flag damage and stay as cool as its glass. Chapter 6′s identity proposal concerns a bodily-affective presentation within an identified subject's conscious life, not a warning symbol or an isolated damage signal. Its advantage would consist in explaining location, felt urgency, and their dissociations through one independently specified profile. If that profile leaves a stable difference in hurting unexplained, the identity loses. Section 10.6 tests whether a further phenomenal ingredient improves the explanation; it cannot derive painfulness from the word *bad*.

In **pain asymbolia**, usually associated with damage to the insular and parietal-opercular 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 at three in the morning is the body, in the dark, insisting (falsely or truly) that something threatens its integrity at that tooth and matters terribly. It was about something all along. It was about you.

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 an intrinsic or projected feature, but then projection must reproduce the same pattern while leaving the private hurt itself without an address. 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 with less machinery.

Pain offers no easy counterexample. Bodily address explains more than a contentless hurt plus projection, while asymbolia supplies a contested but important dissociation.

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, represent affective properties left *unbound*: threat, elation, or bleakness without an object that bears them. That makes her an ally against contentlessness and an opponent of the object-involving reading favored here. Her account honors the felt absence of an object instead of explaining it away. It also asks us to understand property-content without a bearer and owes a further account of why mood so often reorganizes body and field together. The comparison therefore favors a body-and-field account, provisionally: bodily orientation and worldly salience work as two aspects of one situated presentation.[10] The naturalistic story that fixes its content waits for Chapter 15. On that reading, the slot the objection found empty holds the largest object there is. The object was just everything.

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 the bodily weight and the altered possibilities arrive as an organized way of meeting the day. The rival must explain that organization too, but may do so through a contentless cause and its contentful effects. The choice turns on whether we can specify the body's and the situation's contribution independently, then show what would count as getting them right or wrong. A shift in felt mood with that complete presentation genuinely unchanged would count against identity. Chapter 15 owes this application as well as its account of animal perception. Until then, the kitchen gives us a reason to pursue the representational account, not a reason to dismiss the person who finds a feeling we have yet to explain.

Chapter Summary

The claim. Pain and objectless mood form the two strongest families of the intentionality objection. Neither supplies a felt character with no content that would sink the identity claim outright.

Where the argument stands. Pain finds its best single explanation as a bodily threat presented under an evaluative mode; phantom pain shows how its address can go wrong. Asymbolia does not decide between evaluative content and aversive reaction. The body-and-field account of mood remains the view I favor, but a contentless affective background causing changes in presentation remains a serious rival.

What's left open. The account must explain why suffering itself warrants relief, not merely why an alarm becomes costly. Chapter 10 asks whether a further phenomenal ingredient helps; Chapter 15 must explain what fixes the accuracy of pain's bodily address and a mood's diffuse presentation. Neither clinical dissociation nor the kitchen scene settles those questions.

The hand-off. The two strongest cases have supplied no contentless experience. The book now turns to external rivals, starting with Mary in her black-and-white room.

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 figure (the large majority of amputees experience phantom sensations, a substantial fraction of them painful) is standard in the clinical literature; 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, under an evaluative 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 fires 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 discharging its function in the absence of its object. Until that account is in hand, the directedness of pain remains a promissory note, marked here and redeemed in Part Three on the same schedule as the identity claim's own "content of the right kind."

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): 69–89, 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, which is the move Section 7.4 relies on. For the contrasting "tracking" version, on which painfulness is identified with the representation of a specific bodily property, see Brian Cutter and Michael Tye, "Tracking Representationalism and the Painfulness of Pain," *Philosophical Issues* 21 (2011): 90–109.

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 intentional mode, 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. 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 that even vague moods carry representational content in "The Intentionality of Smell" (2014). Angela Mendelovici, "Intentionalism about Moods," *Thought* 2, no. 1 (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. 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; its Chapter 7 concerns thought rather than moods.

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 (the claim that fixing the physical facts fixes all the facts) 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 his premises do not provide. 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. Forty years of 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), 251–271, 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 cogni-

tive-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.

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.

Once experience gets described that way, as inner phenomenal items whose intrinsic properties could in principle obtain independently of any representational relation to the world, no physical process could plausibly produce it. Of course a gap opens. You have stipulated that the two sides occupy entirely different categories. Seal an exhibit behind glass and the unreachability you then discover is the glass. The mystery follows from the description, not from the phenomenon.[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.

What follows from that comes to less than it looks, and the reason sits in how you came by the concept you just used. You know what seeing red is like because you have seen red. That concept arrives by acquaintance and points straight at its object, carrying no description in the middle that reflection could pull apart and lay against a physical description. Its object lies out on the far side of the looking: the red of the tomato, not an inner swatch standing in for it — Section 9.2 left no such swatch to inspect. Acquaintance here names something you underwent, and the concept it installs reaches through that undergoing to the world. Two concepts acquired that differently, one by undergoing something and one by reading, will never lock together on the page, whatever they turn out to pick out in the world. So they come apart in thought, every time, on request.

Chapter 11 makes this its centerpiece. Here it does one job: it explains why the conceiving stays easy no matter how the metaphysics turns out.[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.

Dualism goes wrong in its explanation of these genuine observations. From the fact that experience has a first-person character, dualism infers that experience comes equipped with properties belonging to a separate ontological realm. From the fact that objective description leaves the perspectival out, it infers that the perspectival must be non-physical. From the fact that consciousness matters, it infers that consciousness must be something beyond the physical processes that realize it.

These inferences don't follow. The first-person character of experience is real, but it belongs to a physical organism representing its environment from its particular embodied position in the

world. The perspectival falls outside objective description — not because it belongs to a separate realm, but because objective description abstracts away from particularity by design. That abstracting is its function, its virtue, and its limitation. Consciousness matters because experience engages the world, not because it is made of ghostly stuff.

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 without proving dualism. Moving from a lasting first-person gap to a nonphysical mind adds something to reality when the evidence shows only unequal routes of access.

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. The paper sometimes gets recruited for anti-physicalist purposes it was

never designed to serve. Nagel develops a more nuanced position in *The View from Nowhere* (New York: Oxford University Press, 1986), where the perspectival character of experience comes into sharper focus as a genuine feature of reality compatible with physicalism. 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. The distinction between Levine's epistemic gap and Chalmers' ontological Hard Problem (footnote 3) often gets elided in subsequent literature; keeping it clear matters for evaluating which arguments succeed.

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 diagnostic move here — that the Hard Problem inherits its force from a particular description of phenomenal character rather than from phenomenal character itself — owes its shape to the strong representationalist tradition. 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. 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 conceptual account explains why this conceivability persists regardless of the modal facts; it does not independently establish the identity or refute Chalmers's modal principle. 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 strategy explains the *permanence and form* of the epistemic gap given the identity. 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 same account explains changes in how things look and feel without a further felt basis, while leaving clear what would count against it.

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.

Panpsychism accepts causal closure. It posits no second substance or ghostly intervener. Its 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 the view never claimed a causal gap.

Panpsychism disputes no equation, closure principle, or conservation law. It asks 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 met this diagnosis in print and refused it. Michael Antony pressed precisely this point against him years before the conversion: a sharp concept may track a vague worldly kind.[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. No further armchair datum can break that standoff, since the evidential reach of the armchair is exactly what is contested.

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. That 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.

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 cannot rule out Russellian monism, whose strongest forms avoid the easiest objections. Powerful-quality views join felt quality to causal power in one property, so their hidden basis does real work; phenomenal bonding offers a serious way for many tiny subjects to form one larger subject, though it does not yet explain quality, structure, or the limits of the bond. First-person contact with experience does not reveal a felt basis in basic matter, and no other evidence yet favors that basis over the strong identity between phenomenal character and world-presenting content, an identity learned from experience rather than proved from concepts alone; the book favors identity because it explains more with fewer basic claims and states what would defeat it.

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, no. 3 (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* (Oxford: 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. Tye notes that the no-borderline-cases premise is shared by Jonathan Simon, "The Hard Problem of the Many," *Philosophical Perspectives* 31 (2017): 449–468.

7. Michael V. Antony, "Vagueness and the Metaphysics of Consciousness," *Philosophical Studies* 128, no. 3 (2006): 515–538, and "Are Our Concepts CONSCIOUS STATE and CONSCIOUS CREATURE Vague?," *Erkenntnis* 68, no. 2 (2008): 239–263, argues that our inability to grasp consciousness as a vague-boundaried concept shows a limit on our concepts, not that consciousness is sharp in its own nature — the demonstrative-concept diagnosis pressed in the text. Tye's criticisms of Antony appear in *Vagueness and the Evolution of Consciousness* (2021; cited n. 5), to which the 2024 essay directs the reader ("For criticisms of Antony, see Tye (2021)," n. 4 there); the essay itself restates the refusal in its own voice — "it is part of the nature or essence of phenomenal consciousness that it lacks such cases," something "we don't just believe" but "know a priori" — the sharpness discerned in the phenomenon itself, not read off the instrument; the standoff this produces is described in the text and anatomized in Appendix B.10. 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. 159, "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, no. 3 (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), 121–146, 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 very different routes: through a public account of brains and behavior, and through the direct acquaintance that comes with having the experience. If phenomenal character consists in world-directed content of the right kind, as I have argued, no second felt fact waits for physics to produce or explain. A real limit remains. It concerns our access and our concepts, along with the hard work of finding which physical systems support conscious states; nature itself needs no bridge between two sides of a divide we drew.

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, tested against blur and attention and left open to defeat. **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's own version draws a line most discussions smudge. Give the materialist the metaphysics: suppose felt character consists in some physical condition, and suppose the identity holds. Levine asks not *whether* it holds but whether, granting that it does, we can *see why* it holds. His answer is no. Even with the identity in hand, the felt character does not follow from the physical description the way the liquidity of water follows, once understood, from the behavior of its molecules. There is a step we cannot take in thought. *That* is the gap: not a gap in the world but a gap in our explanation, a place where understanding stalls even after the metaphysics has, by hypothesis, been settled.[2] Chalmers does not infer ontology from that explanatory failure alone. His zombie argument adds a modal bridge. If we can coherently imagine, even after ideal reflection, a world with all the physical facts and no phenomenal fact, Chalmers says that world counts as genuinely possible. 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.

Levine's sentence is true; a physicalist should expect it to be true. 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 hard problem's ontological menace.[4] Refuse that move and Levine's gap remains: real, stubborn, possibly permanent, and metaphysically inert. A true identity leaves precisely this kind of residue because the concepts remain distinct even when their referent is one.

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. No lecture on wavelengths, cone types, and cortical maps will hand a sight-restored adult the recognitional *this*, because only the relevant experience, or one close enough to it, can supply that demonstrative anchor. Mary's Room widens the point into a general account. 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, while the recognitional capacity lets her re-identify the property across later occasions and in other perceivers. The same structure applies to pain, taste, bodily sensation, emotion, and other phenomenal categories. Mary's file on red was complete. She lacked a way of reading a page she already had.

In creatures with our architecture, the acquaintance-grounded and descriptive ways of thinking remain isolated: neither analyzes into the other, and description does not confer the recognitional capacity acquired in acquaintance. No physical description will therefore *yield* the first-person way of thinking. Moving from "such-and-such representational activity" to "*this* — what the coffee is like" does not run by description, because one route describes and the other requires undergoing. One fact reached by two routes leaves a real gap between acquaintance and description. Whether every physically possible knower must share that architecture remains open; Mary* will test it 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, with Tye later rejecting the need for a special class of phenomenal concepts while retaining a materialist role for acquaintance.[8] The book does not need one theorist's proprietary machinery. It needs the narrower structure: description does not confer acquaintance, and a recognitional capacity acquired through undergoing can remain conceptually isolated from a public description of the same physical state. Stated in a line: *the explanatory gap is a fact about our routes to one fact, not evidence for two properties.* The gap becomes metaphysically harmless only if the identity claim holds.

The explanatory proposal goes beyond the observation that a description cannot do the tasting. An occurring state anchors a demonstrative—*this taste*—and later recognition can reuse the capacity established there. Public description reaches the state by its causes, organization, and effects instead. On the house account, neither route supplies an analysis of the other. That architecture can explain why understanding the description leaves a felt demand for something more: the recognitional route remains available to the taster without becoming available through

the description. Whether this accounts for our whole epistemic situation, including the truth and justification of what we learn, is Chalmers's objection in Section 11.7. The acquisition story must not borrow those further achievements from the identity it presupposes.

Section 6.4.3, holding the wide identity claim steady on Inverted Earth, conceded that phenomenal character could change without detection because "the standards of comparison move with them." That consequence supplies no evidence for the wide identity; a structural representationalist rejects it without adding mental paint. Nor does the present appeal to acquaintance buy an infallible gauge back. Acquaintance works in the present tense: token *this* while the coffee does its work and the occurring state anchors the token's reference. The subject may misdescribe or misclassify the state, but the token still picks out the state then and there — the character consisting in the content the state then carries. If the content shifts during the episode, the character shifts with it; local anchoring follows the occurrence rather than preserving an inner sample. Inverted Earth retired the diachronic pretension that this morning's blue can be checked against an inner sample kept faithfully since childhood. That job belongs to memory, whose content gets fixed by the same external traffic as everything else. The recognitional capacity works occasion by occasion. Inverted Earth killed the logbook, not the thermometer; the latter fixes a local reference, not an infallible verdict.

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 dualist's has to be one that 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 — a cost the identity view does not incur.

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 lies elsewhere. On the identity claim, experience sits inside the causal history of every report about it: gap-talk is caused by the very thing it concerns, so judgment and object share a lineage. On the dualist's picture, the processes causing the felt certainty would run identically without the experience. That orphaning invites the debunking. The meta-problem therefore supplies no independent proof of identity, but neither can it become a *tu quoque* against physicalism: judgment and object come apart on the dualist's package, not on the identity view. Chalmers has replies; the sharpest appears below as Objection 1.

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 the identity leaves no further gap to span. It is *explanatorily terminal*: the fact in light of which water's properties get explained, not another fact requiring a bridge. That is where the spade turns.[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, so the identity has something to be terminal *relative to.* 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 has established that felt character and representational activity are one fact reached by two routes, one acquaintance-bound. Section 11.3 then explains why those routes remain conceptually asymmetric. Refuse the identity and the disagreement returns to Chapter 6. The Venus analogy illustrates the moral; the acquaintance-recognition account explains the residual gap once the identity holds.

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 explains how that residue could persist without revealing another property. That does not make our present theory complete: its account of physical realization and of our concepts still needs work. The gap that isn't there, on this view, is the *ontological* gap. Levine's conceptual gap remains.

11.4.1 Where a Conscious State Enters the Story

Deflating the ontological gap does not license leaving state consciousness blank. Chapter 6 now makes a positive, defeasible proposal. A content-bearing state is conscious for a subject when it acquires a recurrently sustained pattern of selective availability and influence across that subject's integrated economy. It can, under suitable conditions, guide attention, working memory, inference, planning, report where report is possible, and flexible action; feedback from those activities can preserve, revise, or displace it. The relevant profile concerns breadth, reuse, recurrence, and two-way correction together, not any behavioral effect whatever.

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": they leave wider and more durable effects on what the system remembers, says, reasons about, and does.[13a] That picture gives *poising* a tractable meaning. Poise is no secret doorway through which a content passes and lights come on. It names a content's place in a distributed competition for access and control.

The account leaves room for genuinely graded cases. A content can gain enough reach to guide one flexible action without surviving into memory, or recur locally without entering wider planning. No point-sized moment of presentation follows. Global-workspace, recurrent-processing, predictive, and higher-order theories now propose ways of realizing aspects of the profile rather than supply a menu in place of the book's answer. A matched conscious state outside the profile, or an unconscious state matching it in content-sensitive breadth, recurrence, flexible control, and correction, would defeat that answer. The identity of phenomenal character with the complete world-presenting profile would remain a separate claim, exposed to 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 offered an inference to the best explanation, showed how blur and attention constrain the proposed profile, and stated what would defeat it. It also granted that necessary grounding can match the evidence without mental paint. The present chapter cannot convert that remaining theory choice into an empirical victory merely because identity would retire a difficult question.

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.

One further debt remains open. The identity claim invokes representational content "of the right kind," and Chapter 6 separated the phrase into several questions. Part Three develops the **content-fixing** account: teleosemantics explains in causal-functional terms how selection, learning, tracking, producer-consumer organization, and world-involving use can make a state answerable to what it concerns.[17] Its affective applications remain incomplete: the content of pain's badness and diffuse moods, and our warrant for seeking relief from pain itself, still need fuller accounts. Content fixing also leaves phenomenal mode, the distributed consciousness profile, and the boundary of the subject as separate questions. Those remain with Chapter 6′s empirical proposal, Chapter 17′s account of the persisting control organization, and Part Four's bearer-sensitive tests. The deflation therefore leans forward as well as back: back on Chapter 6 for the identity and state-consciousness profile, forward to Chapter 15 for the account of aboutness that keeps the identity from resting on unexplained content.

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 that second horn. Let C mean that an occurring state anchors a demonstrative—*this*—and a capacity to recognize similar states later. A physical duplicate shares that architecture. It explains why the two concepts remain isolated and why both Mary and her zombie would talk

about a gap. It does not establish that Mary's thought is true. Chapter 6′s identity claim does that work: if phenomenal character consists in the physical world-presenting state, Mary's thought reaches its truthmaker. The strategy therefore explains the gap if identity holds; it cannot support the identity without circularity.

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]

The answer begins with the difference between a property and its machinery. Acquaintance reveals part of the state's nature: the organized way a red surface or bodily disturbance is presented here and now. It does not reveal the neural and historical machinery that realizes that profile. On Chapter 6′s identity, acquaintance thereby reveals something physical without classifying it as physical. Present pain secures that pain occurs and how it presents the body; it does not disclose a separate intrinsic constituent or the property's complete metaphysical nature.

Helen Yetter-Chappell presses from the opposite direction. Her Mary* has complete physical knowledge and genuine phenomenal concepts, but her cognitive architecture automatically matches the information carried by a recognitional state to the physical description of that state. If physicalism holds, Mary* can apparently know the phenomenal-physical identity *a priori*. The case shows that physicalism need not impose our conceptual architecture on every possible knower. It does not show that our gap marks a second property. The book's conclusion stays exactly where the two objections leave it: acquaintance explains our route; identity fixes what the route reaches.

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.

What the chapter can claim is more modest. Introspection secures the felt character and its first-person presentation; it does not, without further argument, secure the character's intrinsicness. The identity account explains both the datum and the stubborn sense of residue through the split between acquaintance and description. The rival preserves the residue as a further property but inherits the causal and explanatory costs already counted. No phenomenon gets eliminated. What loses its automatic authority is the inference from *there is undeniably something it is like* to *therefore an intrinsic constituent 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.

The reply begins by noticing the argument's form. It runs from explanatory failure to framework inadequacy: from *our best physical accounts do not make consciousness intelligible* to *the natural order must contain something those accounts leave out.* That is the epistemic-to-ontological slide this chapter has been diagnosing since its first page, executed at the scale of a worldview rather than a concept-pair, and it inherits the same unpaid debt — some reason to treat intelligibility-to-us as a constraint on what nature contains. Section 11.4 said why an identity refuses exactly that demand. If felt character consists in poised, world-directed content, no further fact remains for a successor science to render intelligible; the demand that a new framework explain the appearance of materially irreducible minds only finds work to do once the slide has already been granted. The teleological principles wait on the same grant. They would have explanatory purchase only if minds arrived as an addition to nature's furniture; given the identity, nature grew minds the way it grew metabolisms, by ordinary selection among physical arrangements, and a new kind of law would find the job already done. What survives is a conditional the book has already signed: if the

identity claim falls, the scale of Nagel's response stops looking extravagant — Section 11.6 names that hostage openly. The live disagreement concerns not honesty about the gap but whether an a posteriori identity makes a legitimate place for explanation to stop. Chapter 6 argued that it does. *Mind and Cosmos* has arguments, and they are not the weak ones its critics reach for first. What it does not have is a reason to deny that such an identity can be a terminus — and without that, the framework case never reaches the identity claim to knock it over.

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. The two concepts remain epistemically independent: one gets acquired by undergoing and recognizing the state, the other through public description. No description, however complete, by itself installs the recognitional capacity it describes. A perfect account of facial geometry without perceptual exposure to Gordon does not hand you the ordinary ability to recognize him in a crowd, and a perfect account of wine chemistry does not put the vintage on your tongue. In those cases the failure of derivation reveals something about concept acquisition, not a second non-physical face or wine. The phenomenal case presses harder because acquaintance supplies the only route to the concept, but the form of the asymmetry remains the same.

Chalmers can still insist that a necessary identity without an a priori route from one side to the other would stand alone in nature. Type-B physicalism accepts that revisionary consequence. The acceptance is not cost-free or licensed merely by loyalty to physicalism. Chapter 6 supplied the independent case for the identity; Chapters 9 and 10 counted the causal and combination costs of its rivals. Strong necessity is the revisionary item this view carries after that comparative accounting, not a device introduced here solely to rescue it. The exact point of refusal is Chalmers's restricted principle that ideal primary positive conceivability entails primary possibility. Ideal reflection perfects reasoning with the concepts available; it need not erase the isolation between a concept acquired in acquaintance and one acquired in description. Nothing in the modal argument

independently ranks Chalmers's bridge principle above Chapter 6′s identity claim; each yields an opposing verdict about the same scenario. That leaves a local standoff about modal method, not a neutral verdict on the book's case. The book sides with the independently argued identity and treats ideal conceivability as defeasible evidence. Its price is strong necessity: the identity remains necessary although no a priori route joins its terms.

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, these changing, world-directed presentations constitute the felt life whose place in nature we wanted to understand. 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 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. The stronger proposal concerns our concepts: acquaintance grounds a way of recognizing the state that public description does not supply. This explains the persistent gap if the identity holds; it neither establishes that identity nor defeats a necessary-grounding rival that matches Chapter 6′s evidence.

What's left open. We must still ask why nature contains conscious, world-directed organization and how different physical systems realize the distributed access-and-control profile. Borderline cases and subject boundaries remain empirical. The authority of ideal conceivability as a guide to possibility also remains in dispute.

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. The distinction governs the whole book and is drawn already in Chapter 2 (n. 3 there): Levine separates the metaphysical question (are phenomenal properties identical to or constituted by physical properties?) from the epistemic question (does any physical description make intelligible *why* a given physical state produces *that* felt quality?), and argues that even if the metaphysical answer is yes, the epistemic gap may persist. The position developed across Chapter 9 and the present chapter is precisely that Levine's gap stays open while Chalmers's slide from explanatory failure to non-supervenience does not go through.

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 three-step gloss — "not made intelligible by" → "not necessitated by" → "not identical to" — compresses the structure of the anti-physicalist argument from the explanatory gap. The first step is Levine's and is granted. The second and third are the disputed escalations: from an epistemic failure of *a priori* entailment to a modal claim of non-necessitation, and thence to the denial of identity. The book's claim is that the epistemic premise is true and the modal/ontological conclusion does not follow, because *a posteriori* identities standardly exhibit exactly this profile — true, yet not knowable by reflection on the concepts alone. See n. 11.

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 book's reply to that wrinkle is the phenomenal concept strategy of Section 11.3, which locates the appearance/reality slack in the *concepts* rather than denying the identity. 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 the *explanatory* gap an a posteriori identity leaves behind is no objection to its holding.

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).

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); 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; the program itself 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), 141–205, 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 explains why the scenario remains ideally conceivable if the identity holds; it does not refute Chalmers's modal principle or establish the identity. For type-B replies, see Block and Stalnaker (cited n. 11), Brian Loar, "Phenomenal States" (cited Chapter 8, n. 12), 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.

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.

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 the zombie's corresponding beliefs can differ in truth, justification, or cognitive significance. The text therefore takes the second horn and limits C's work. Helen Yetter-Chappell, "Dissolving Type-B Physicalism," *Philosophical Perspectives* 31 (2017): 469–498, esp. secs. 2.2–2.4 and 3.2, doi:10.1111/phpe.12099, constructs Mary* with complete general physical knowledge, phenomenal concepts, and an architecture that automatically matches recognitional information to physical description. She argues that Mary* can know the identity *a priori*. The text accepts the architectural pressure while resisting any inference from it to a second property.

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 accepts limited translucency about the manifest world- or body-presenting profile and denies that acquaintance reveals a complete categorical essence or a second phenomenal aspect. 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), esp. chs. 3–4. Nagel's skepticism is explicitly not religious in motivation, and he endorses no definite alternative; the constructive suggestion runs to principles of natural order teleological rather than mechanistic in logical form.

The 1974 bat paper stopped short of the metaphysical verdict — it closes on our lack of concepts, not on what exists, a restraint the present book reads as proto-diagnostic (see Chapter 2, n. 1, and Chapter 9, n. 1) — and *Mind and Cosmos* does not stop short, which is what makes it a framework-level challenge rather than a restatement of "What Is It Like to Be a Bat?", *Philosophical Review* 83 (1974): 435–450.

21. The contrast with the rivals is exact and was earned in the preceding chapters: substance dualism (Chapter 9) adds a second realm; panpsychism and Russellian monism (Chapter 10) add fundamental phenomenal character at the base of the physical world. Both answer the hard problem by granting its central demand — that consciousness needs *more* than the natural-relational world to be what it is. The book's deflation refuses the demand: consciousness needs no addition, being a high-level relational achievement of physical systems (the identity claim), and is on that account the more remarkable for requiring nothing extra. 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

> *"Our general framework of particular-reference is a unified spatio-temporal system of one temporal and three spatial dimensions."*
>
> — P. F. Strawson, *Individuals*

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 causal-historical route 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. The singular/general contrast is Evans's (*Varieties of Reference*, chs. 4–6): a general thought represents a kind and underdetermines which particular falls under it, whereas a singular thought reaches a determinate object. 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. Benacerraf's familiar identification problem explains the remaining qualification: 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* (Oxford:

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 applies Evans's singular/general distinction (*Varieties of Reference*, chs. 4–5; see note 6) to 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 builds the relation into a description — reference fixed by 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. The revised argument grants the vocabulary. The expression may remain descriptive; its success depends on the actual token and actual causal ancestry that make its reflexive clause true. Causal descriptivism preserves descriptivist semantics by accepting the relational metasemantics this chapter requires.

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.

> Imagine looking up an unfamiliar word in a Chinese-Chinese dictionary. You find the entry, but the definition consists entirely of other Chinese characters you don't know. You look those up. More characters. You look those up too. You could spend forever in the dictionary without making contact with anything the characters mean, because meaning, Harnad argued, cannot arise from symbols defined solely in terms of other symbols.[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. Measured Shannon's way, each page against the statistics of its source, a page of random noise carries more bits than a page of Shakespeare. English runs predictable and noise does not, and a bit measures exactly that. Nor

does the measure attach to the page alone. Its number falls out of the ensemble the page was drawn from; change the ensemble and you change the number. It still means nothing at all.

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. One lawful correlation will bear a dozen readings. The state that covaries with a fever covaries just as lawfully with the infection that caused the fever, the inflammation that drives it, the thermometer reading that records it, and the worried frown that follows it. Those reliable correlates run outward in both directions along the causal line. Nothing *in the correlation* 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: Oxford University 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. Genuine representation requires a state that can be *false* — that has a determinate content it can fail to match — which is why the bare information relation is insufficient for aboutness. 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): article 12307 (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 — the claim that bears hardest on the account developed here — that multimodality and embodiment are *neither necessary nor sufficient* for referential grounding. That last point denies a premise this book leans on. Chapter 23 answers it against the live systems, and Chapter 24 answers the referential-grounding horn directly.

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: Oxford University 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. The point recruited here is narrow: a formal symbol system exhibits none of the directedness Brentano made criterial, so formal symbol manipulation alone cannot constitute a mental state.

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, and the animal may gauge prey distance in ways that look faintly constancy-shaped. Ruth Millikan anchors representation in the evolutionary function of just such mechanisms.[6] Burge asks whether the animal perceptually objectifies a stable distal thing; 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 nevertheless give its states a thin, subpersonal content, while the animal's wider organization determines what perceptual and cognitive capacities those states help realize. This chapter withholds *perceptual objectification*; the next earns the thinner representational claim and, with it, the ability to miss a fly.

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.

The frog may still fall short of Burgean perception if its prey-capture circuit never treats a fly as a distal object separable from its retinal trace. Higher vertebrates, including human infants quite early in development, achieve that much. They see the cat, not the cat-projection. Burge's and Millikan's thresholds answer different questions; a creature may clear the functional-semantic threshold before it clears the perceptual-objectification one.

14.1.3 A Text-Only Model Has No Constancies to Solve

Bring this back to large language models. Tokens arrive already objectified: 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 constancies that hold a perceived property fixed through changing stimulation. A text-only model therefore achieves no Burgean perceptual objectification, though its mechanisms may carry thinner, task-shaped structure.

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 thermostat fails this distal test. The frog may carry subpersonal content without meeting Burge's full threshold. A text-only model has no perceptual channel through which a scene could impose constancies to recover. Chapter 15 asks whether selection and causal use can nevertheless give a mechanism correctness conditions.

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. The proposal has causal substance and may establish content genuinely answerable to linguistic distributions. Whether that content counts as original remains unsettled.

Mallory clears the causal bar: intervene on the vectors and prediction changes; alter the loss and training shapes a different organization. Linguistic-context structure does explanatory work inside the mechanism.

But causal privilege and semantic independence are different achievements. Public practice supplies the word-types and meaningful uses in the corpus; the training design supplies the context window and loss. Those facts do not show that practice exhausts the learned state's correctness standard. How often a word occurs in a context is also a fact about the world. Training may make the state answer to that fact rather than merely to an interpreter's preferred reading.

Coordinated relabeling cannot decide the issue. Permute the word indices and corresponding matrices while preserving their links to actual word-types: the code changes, but the word-to-context relation need not. Change the external pairing instead, and a relation that may help fix content has changed. Geometry alone cannot choose its interpretation; Mallory's argument appeals to history and use as well.

Human learners inherit public standards too, so manufacture marks no dividing line. A child can learn facts about word use as well as facts about trees. Neither subject matter settles originality. For either learner, the question asks whether history and use fix accuracy in a way not exhausted by interpreting practice. Word2vec's limited case remains open on that test; later perception, testimony, correction, and action may support further content.

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 demon-

strative 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?

Natural selection, individual learning, culture, and artificial training can all contribute to the second relation. A system may borrow its vocabulary while continuing traffic with the world fixes new uses and errors; another may internalize a human standard that drives prediction from within while remaining derived. Originality can therefore vary by content and grain, though each attribution must answer a definite question about its standard. Mallory's embeddings clear the causal bar. Their independence at the grain of linguistic distributions remains unsettled; inherited vocabulary alone neither grants nor defeats it.

Dennett offers the strongest reason to discard the distinction. On his real-pattern view, intentional description earns literal standing where it yields robust, economical prediction across novel and counterfactual conditions; no semantic spark or first interpreter needs to enter the story.[14] The book agrees that intentionality can emerge gradually and that objective patterns need no ghost behind them. It retains one further diagnostic. A pattern may support prediction while the accuracy standard attributed to a state still comes entirely from our practice; a producer-consumer mapping may instead make the state answerable to a world independently of what an interpreter

calls it. That difference concerns semantic dependence, not whether the content is real and not whether the system has crossed a general gate into mindedness. Dennett may treat the difference as one more feature of the pattern. This book gives it a name because it changes what could correct the state when our interpretations fall away.

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. Burgean objectification distinguishes perception from registration. Teleosemantics may give a frog's subpersonal circuit original content before the animal clearly meets Burge's perceptual threshold. Mallory shows how word2vec can carry causally potent content about worldly linguistic distributions; originality at that limited grain remains unsettled. Neither result turns a circuit into a perceiver or a model into an understanding speaker.

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.

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 is central to Millikan's account and is what distinguishes teleosemantics from mere informational semantics à la Dretske (*Knowledge and the Flow of Information* [Cambridge, MA: MIT Press, 1981]). Information by itself does not generate the normative dimension — the difference between succeeding and failing at representation — because mere correlation between sign and signified does not yet involve the kind of cooperative history that lets a sign be *wrong*. The bee-dance example is Millikan's own (*Language, Thought, and Other Biological Categories*, 96–98); the philosophical point is that proper function lives downstream of consumer uptake, not in the producer's intrinsic states. 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 that learned content may

genuinely answer to worldly linguistic distributions and leaves originality at that grain unsettled. A definite derived verdict would require showing that interpreting practice exhausts the distribution-directed standard; a definite original verdict would require establishing the relevant independence. 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.

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, while Swampman, a perfect copy with no past, shows why present form may fall short. 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. It shows that selection history cannot be the whole story. Evolution explains why states exist and can fail; task functions, tracking relations, and consumer use must help fix their content.

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 lies below every discrimination the wider organization can use, and therefore below any phenomenal difference on the identity claim. Mechanism-level content supplies the semantic ingredient; perceptual mode, integration, and poising locate the larger world-presenting state.

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. If that description preserves every consumer-relevant counterfactual, content has not earned its keep.

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 needs its own constraint. Why does the state present a threat as bad rather than merely locate it? On the proposed account, its history and use must distinguish conditions that impair bodily functioning from harmless changes, and assign them different priorities across attention and protection. A mere increase in avoidance would not establish that content. The account needs evidence that this mapping explains the organized response better than a neutral damage signal followed by aversion.

Mood makes the unfinished work more visible. For Chapter 7′s body-and-field account, candidates include bodily depletion and a surrounding situation offering few manageable possibilities for action. Their accuracy would depend on the body and available possibilities, not on how convincingly the day feels bleak. Which candidate a mood represents requires evidence about its history, use, and correction. The general theory supplies a test; it does not yet fix every mood's diffuse

content. That remains a substantive limit on the promised naturalization, not a reason to make felt character fundamental.

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. Chapter 14 grants the embedding mechanisms causally active content that may genuinely answer to worldly linguistic distributions. Whether that

content counts as original remains unsettled. Either way, the semantic finding does not settle the level of attribution. Even 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.

15.5 What Embodiment Contributes

Descartes treated the body as incidental to the mind. The *res cogitans*, the thinking thing, is pure intellect attached to a body that brings hunger, fatigue, and misperception. Its essence remains independent, and in principle it could exist disembodied.

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.

Under strong representationalism, a system with no content of the right perceptual kind has no phenomenal character. At $t = 0$ there is nothing it is like to be Swampman, a phenomenal externalism Dretske accepted explicitly.[13] This follows from the account; it supplies no independent evidence for it. The internalist can instead make phenomenal content supervene on present organization, but that gives up the book's wide answer to Inverted Earth. I accept the other horn: complete phenomenal content includes the missing history. Millikan and Neander resisted the common Swampman verdict with open eyes; intuition does not decide the issue.[14]

The comparison with a trained language model now concerns history, not manufacture. Swampman lacks a content-fixing history. A text-trained model has a training history that may make its submechanisms answer to actual linguistic distributions. Those distributions belong to the world, even though public practice produced the words. Originality at that grain remains unsettled; inherited materials do not decide it. Model-level understanding remains open too. 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. Chapter 14′s narrower verdict on embeddings carries over: content may genuinely answer to linguistic distributions, while originality at that grain remains unsettled. Neither conclusion establishes whole-system understanding or distal reference beyond the history that fixes the content.

What's left open. The general model still owes fuller accounts of pain's evaluative content, mood's diffuse accuracy conditions, and why pain itself warrants relief. Integration can make content a reason; reasons-responsive control and moral answerability require more. A state's perceptual mode and conscious role also remain distinct from the organization that bears it.

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 and 7–8. 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.

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," *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.

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 the same wide reading against the charge of absurdity 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.

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,'" 143

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.

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 310 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.

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. Even Kant's formal condition for unified experience names no tenant. 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 useful 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*. We can make our past character more determinate from the present, settling by a creative step matters that had no answer before. Perhaps the person you became settles whether the young you was already brave or already leaving. Real pasts, Dennett argues, do not take late edits. **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.

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 could no longer 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 remains a separate question about autonomy, interests, and welfare, 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; the behavior is dose-dependent and typically remits when the agonist is reduced or withdrawn. The vignette composites the now-standard clinical picture. For the causal claim see Melissa L. 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 mechanism is read as agonist overstimulation of mesolimbic reward circuitry mediating impulse regulation, with the patient's evaluative knowledge preserved — sufferers routinely conceal the behavior and describe it as against their own better 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. The philosophical use made of the case here requires only the clinical sequence itself, which no one disputes: cause installed, capacity lost, cause removed, capacity restored.

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 treat the causal completeness of the neural story as the discovery that settles the question, whereas the professional literature has for over a century declined to treat the bare fact that choices have causes as settling anything — libertarians denying the 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 moderate reasons-responsiveness of the mechanism that issues in action, 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 is standardly a current-time-slice property of the mechanism; Fischer and Ravizza add a distinct *historical* condition — the agent's "taking responsibility" for the mechanism over time. 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* (New York: 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*

CHAPTER OVERVIEW

Grief can show in a friend's face before a word gets said. I do not think we first inspect bare moves, compare them 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. Remove the room, and concern, pain, and joy can appear in a life joined with ours, though we can misread them and must often yield to the other person's word. Smooth speech at a chat window can spark the same sense of presence without proving that one lasting subject made it. I therefore ask what our sense of another mind tracks: speech and action linked with sight, memory, error, and change across one shared life.

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 metaphysical chasm dissolves. What remains is

harder, more ordinary, and more interesting: fallible knowledge of people whose lives become available in engagement without ever becoming ours.

19.1 The Argument from Analogy

> Take the argument seriously. 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.

Objection: the science says 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. **Reply.** 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 can hold for concern. In difficult cases I may reason explicitly, gather testimony, or revise a first impression. Ordinary recognition can still count as perceptual in the important sense: concern is what the expression presents, subject to defeat. Science may explain how recognition works without turning its warrant into Mill's argument.

19.3 Fallible Knowledge of Other People

The metaphysical chasm dissolves. The work of understanding particular people remains.

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 correction can defeat an initial reading, just as later light can reveal that the gray shirt was green. 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 chasm 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, the person stands present in the expressive, embodied life one encounters — present in it, not inferred behind it. 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 few lines of pattern-matching 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. An artificial system whose content-bearing states entered a persistent, world-sensitive organization of memory, correction, planning, and action would qualify for consideration on the same terms. 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. Analogy can help us understand another person, but it does not supply the basis of every encounter. Treating it as our only route to other minds assumes the inner-theater division between a public body and a hidden mental life.

Where the argument stands. First-person authority remains distinctive but defeasible; it discloses no private inner object. In ordinary second-person life, expression can present concern directly without becoming infallible or exhaustive. Acting, masking, and error defeat particular readings without turning every recognition into Millian inference. The metaphysical chasm dissolves; the fallible work of understanding people remains.

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's "Criteria, Defeasibility, and Knowledge," *Proceedings of the British Academy* 68 (1982): 455–479, supplies an important alternative: psychological criteria can provide defeasible grounds for knowledge without first being translated into psychologically neutral evidence. Defeasible need not mean derivative.

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. 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. As the later Putnam realized, his cheese was only half the story. He left the world out.

Start with a program too simple to confuse anyone. Almost every programmer begins with some version of this one, here in C++:[1]

```
#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 written those words on a piece of paper. The paper does not think or speak. Nobody is home. This establishes an easy baseline: producing meaningful words does not by itself establish a mind.

Now add memory, recursion, and billions or trillions of parameters calculated from machine learning on enough text to fill ten Libraries of Congress.[2] At some point, according to some computational functionalists, the program 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. Size alone supplies neither the claim nor its answer.

Philosophers call the broader theory *functionalism*, and for several decades it came as close to orthodoxy as philosophy of mind allows.[3] 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.

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. Pain is whatever state gets caused by bodily damage, 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.

From the late 1960s, David Lewis turned the picture into a method. 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 joined two claims. Material anti-chauvinism denies any substance a monopoly on mentality; actual recurrence says a mental kind does occur across disparate realizers. Keep the first and treat the second 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. An abstract program specifies formal states and transitions among them. **Strong computationalism** identifies those computational states with mental states. 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.

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.[10] 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.[11] 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 Three Tests of the Wide View

A critic may now suspect that wide functionalism has won by moving the boundary whenever trouble appears. Block's homunculus-nation, the relabeling objection, and Chalmers's replacement argument test whether the widened view says anything sharper than "include whatever matters."

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.[12] 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.[13] 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.[14] 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 alone cannot settle the case. Exotic composition may mislead us; if the stipulated facts truly sufficed for consciousness, disbelief would not make them stop sufficing. Wide functionalism offers a diagnosis rather than a louder intuition. 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. The wide view should grant every one of those causal facts. What the stipulation does not settle is which contents those transitions carry for the assembled whole, whether any state of the whole presents that content under a bodily-affective mode, or whether the activity belongs to a persisting subject with the required attention, poising, and integration.[15] Present contact is not by itself a representational past.

An ahistorical simulator may match the transitions without an independent basis for their contents. A copied mechanism inserted into a content-bearing perception-action loop may inherit the roles its histories and consumers already fix; a persisting nation could acquire an equivalent wide organization through its own encounters. The narrow machine table cannot tell us which case we have.

This answer must accept its own consequence. Suppose the nation-cum-body persisted, learned through its own encounters, developed wide states whose consumers answered to damage and color, integrated those states through memory, inference, planning, and action, and realized the perceptual and bodily-affective organization required for consciousness.[16][17] Then the material of citizens and radios would supply no objection. Wide functionalism should count the system as a candidate mind. If it still said no merely because a nation feels like the wrong kind of thing, it would relapse into the chauvinism it rejected in Section 20.1.

Block's one-hour case exposes how much narrow duplication leaves open. It does not prove that the nation lacks content or consciousness; it shows that the stipulation settles neither.

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 explanations. 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. I am content with that result. The chapter criticizes a substantive thesis, not a word: the claim that internal causal or formal organization suffices. Anyone who rejects that claim, refuses material chauvinism, and explains mentality through the right world-involving organization has crossed the important distance, whatever appears on the folder tab.

20.4.3 Organizational Invariance

> The strongest test comes from David Chalmers, a property dualist who nonetheless defends the *principle of organizational invariance*: systems with the same fine-grained functional organization have qualitatively identical experiences.[18][19] 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 the replacement verdict while rejecting Chalmers's universal principle as stated. The replacement 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. The process holds the subject's complete world-presenting organization fixed while changing the inner material.

The case shows that material substitution can leave mentality unchanged when the subject's wide organization and history remain fixed. It establishes one route across materials, not a general multiple-realizability thesis. Chalmers does not show that world and history can be dropped from the type being preserved.[20]

Separate three cases.

1. A replacement descendant inherits the original subject's content-fixing history and complete phenomenal organization.
2. A freshly assembled narrow duplicate matches the internal topology, judgments, and utterances but does not receive their broad content merely from that match.
3. A differently built system may independently acquire an equivalent history and wide organization.

Wide functionalism grants the relevant capacity in the first and third cases. The narrow description leaves the second unsettled. During gradual replacement, judgments retain their worldly content because the subject's history stays fixed; in a fresh duplicate, matching transitions and sounds do not establish the same content or phenomenal mode.

If Block's system acquired an equivalent history and integration, the difference would narrow; if every wide relation relevant to the attributed capacity matched, 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.

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, which seems to require another rule, and 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.[21]

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.[22] 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.[23] 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.[24]

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, though material replacement can preserve a state when the subject's wider organization and history remain fixed.

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. 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.

2. The premise commands near-universal assent, though not always for the same reasons. Panpsychism and Russellian monism grant it because on their account nothing phenomenal comes into being anywhere, phenomenal properties sitting at the fundamental level rather than arriving with any particular arrangement of matter; 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.

3. 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 the further claim that the relevant functional states are computational states. 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 Founda-*

tions 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.

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. *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, esp. secs. 7–9, 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. 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.

11. 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.

12. 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.

13. 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.

14. 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 the *Stanford Encyclopedia of Philosophy*, "Qualia," sec. 4.

15. 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.

16. 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.

17. 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.

18. 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.

19. 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, which Chalmers takes to be incoherent.

20. The chapter rejects Chalmers's universal principle as stated. 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: a replacement descendant inherits broad content because its history remains fixed. A freshly assembled narrow duplicate does not acquire that content merely by matching internal topology, judgments, and reports. 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.

21. 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.

22. 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. II–VI. 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.

23. 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.

24. 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. The change matters: training can give internal mechanisms causally internalized derived content, and it yields fluency, abstraction, code, and transfer no hand-written room could approach. It does not by itself answer the Room's question. 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 closes the shortcut from syntax to understanding but leaves the systems reply open. 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 that computation is observer-relative while the brain's causal processes are intrinsic is the basis of Searle's claim that "the brain is a causal machine and not a computational machine." This is a strong claim and has been contested; for a defense of the observer-relativity claim see Searle, *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992), ch. 10, 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. 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."*
>
> — Alfred Korzybski, *Science and Sanity*

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. Chalmers's counterfactual repair 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. A rock does *not* implement every finite-state automaton, because its physics carries none of the rich, counterfactual-supporting, component-wise structure genuine implementation demands.[5] Putnam's mapping laid a story over the system instead of finding one in it. Implementation, on this account, is a substantive constraint. Most systems implement rather few computations. 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. Now build me 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. Deutsch's physical principle says that every finitely realizable physical system can be perfectly simulated by a universal model computing machine operating by finite means.[10][11] Neither claim says that computing a model of a process instantiates every causal power of the process modeled. A brain may be perfectly simulable without its simulation thereby possessing the powers that make a mind.

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.

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. "Some people will say this doesn't count for what's needed for grounding because the environments are virtual," he writes, and answers: "I don't agree. … I've argued that virtual reality is just as legitimate and real as physical reality for all kinds of purposes."[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*, treats training through the contextual and inferential relations of human language as enough to ground minimal semantic contents and bounded but genuine understanding. On that account, acquiring and using the right dense, context-sensitive relations among symbols realizes understanding rather than merely describing it. A data center need not move air to achieve that.

Training can give internal mechanisms rich causal structure and genuine content about linguistic contexts. When several content-guided capacities correct and constrain one another, the book also grants a basis for partial understanding. Havlík and I therefore need not disagree about every bounded achievement. The remaining questions concern what fixes the content and how broadly the larger system integrates its use. Neither a demand for adult human breadth nor the hurricane can answer those questions for us.

But I do not want to rest the reply on "you assumed your conclusion," because the fragmentist can fairly say the same of me. An independent reason limits the portable web—a theorem's worth of reason, not a clash of starting points. A system of relations fixes its relata only *up to structure*. Give

me any web of symbol-to-symbol relations, however dense, and I can relabel every node, and so long as the swap preserves the pattern of relations, the new assignment satisfies the web exactly as well as the old. The relations cannot tell the two assignments apart, because the relations are all there is. M. H. A. Newman raised this objection in 1928 against a purely structural account of knowledge, and Putnam made it the decisive step of his model-theoretic argument: no amount of internal or structural constraint, by itself, pins a term to its referent.[18] A distributional web inherits the predicament. Its geometry alone cannot select which worldly thing each symbol is about. Training may supply genuine content about linguistic contexts; a further causal-historical chain may also fix distal reference. The theorem blocks structure *alone*, not those historical routes.

Nor does richness buy an exemption. Probe a trained network and you can find internal structure that genuinely mirrors a domain—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. Grant that training made the geometry a network achievement and that the mechanism uses it causally. Whether the resulting content is original still turns on what fixes the task's semantic standard. The permutation shows that geometry alone does not settle that question or distal worldly reference. Nor does mechanism content by itself establish understanding by the larger system.

The result is multiplicity: every relation-preserving relabeling satisfies the web equally well, so the relations select no distal reading. Reference remains possible if something beyond the geometry breaks the tie — selection history, corpus-mediated causal chains, or live world traffic. Co-occurrence alone gives us more symbols: the merry-go-round, spun faster.[19]

A reader who knows the literature will have a reply ready, and it deserves stating at full strength. David Lewis answered the permutation argument a generation ago: reference gets fixed by 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. Grant that the world's natural joints break some structural ties. That may help fix reference, especially when selection history and causal traffic engage those properties. It still does not, by itself, turn a self-enclosed web into an understanding subject; whole-system attribution remains a further question. The hurricane survives intact.

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 fix original content only where correctness standards are not exhausted by interpretation; derived content remains real, and neither kind establishes whole-system understanding.

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 that intrinsic realization without remainder. 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 welfare 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 realizable 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. The principle remains contested in its strongest form; the chapter grants it precisely to show that even at full strength it licenses nothing the brain-simulation hypothesis needs.

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. The identity claim this book defends at Chapter 6 holds in the main text that phenomenal character *consists in* representational content of the right kind; the "right kind" is cashed out by the world-involving, teleosemantic relations of Part Three (Chapters 14–16) — exactly the relations a merely formal simulation describes but does not thereby instantiate.

13. The homunculus-nation and the absent-qualia objection are built and sourced in Chapter 20, Section 20.4.1 (notes 12–17); 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. Functional reproduction secures the first by construction and leaves the second untouched — which is exactly the leverage the brain-simulation hypothesis lacks.

15. David J. Chalmers, "Could a Large Language Model Be Conscious?," *Boston Review*, August 9, 2023 (from a NeurIPS 2022 address); the quoted sentences appear in his 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 welfare.

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), esp. secs. 3.5–3.6, https://doi.org/10.1007/s11229-024-04878-4. 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. The view is the most serious recent attempt to argue that understanding, unlike weather, is the kind of phenomenon a formal system can realize — which is why it, rather than the boosters' enthusiasm, is the objection the chapter answers at strength.

18. That a system of relations fixes its relata only up to structure—leaving distal reference undetermined by structure alone—is 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 the decisive step in Hilary Putnam's model-theoretic argument (*Reason, Truth and History* [Cambridge: Cambridge University Press, 1981], ch. 2 and the Appendix). A distributional or fragmentist web inherits that limited result: its geometry alone cannot fix which worldly thing a symbol is about. Selection history may still privilege an encoding, and causal-historical chains may help fix distal reference. Chapter 23, note 8, brings the same distinction to large language models. Interventionist probing has recovered rich, behavior-controlling internal structure from sequence models—the model case is Kenneth Li et al., "Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task," ICLR 2023 (arXiv:2210.13382). Training can make that geometry causally operative and content-bearing. Where an interpreting practice exhausts the correctness standard, the content remains derived; inheritance of the task alone does not establish that dependence. The permutation result limits internal geometry alone, not an account that also appeals to actual content-fixing history and use. Neither geometry nor mechanism content by itself establishes whole-system understanding.

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: a web of intra-symbolic relations does not *by itself* fix distal worldly reference, though training history and corpus-mediated causal chains may add what bare geometry lacks.

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, with a test that pairs the same inner pattern with different worldly things, shows why form alone cannot fix what words point to. Training can still make inherited content guide a model from within. A body of text, a camera, a tool, or a long chain of feedback may also give its states a real path to the world. Neither gain by itself makes the whole model an understander. I count the whole as an understander only when several content-guided skills check and shape one another through memory, learning, plans, and action in one lasting system that can find out when the world has proved it wrong.

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.

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]

> 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. What listening alone cannot provide is 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; it has not thereby connected "stick" to a stick or "bear" to a 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.)

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 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 may establish content about linguistic distributions, with originality at that grain still open. 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, so a word's distribution, the full profile of contexts it keeps, can stand in for its meaning. That hypothesis underwrites every system this chapter is about.

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 intervention-stable effects on later processing. Fintan Mallory's consumer-based teleosemantics treats those facts as sufficient for genuine, original content about linguistic contexts.[7] He has earned the causal conclusion. The learned mapping operates inside the mechanism and helps explain how training shaped the parameters.

The semantic question remains open at Mallory's proposed grain. Linguistic distributions belong to the world: how often *cat* occurs in a context need not match anyone's preferred account of its use. Training may make a state genuinely answer to such facts. Public practice supplies the words and corpus, but that ancestry does not show that interpreting practice exhausts the learned standard. A coordinated permutation of indices and matrices may preserve the relation to actual word-types. Changing the external pairing instead changes a candidate content-fixing relation. Neither operation decides whether the content is original.

The frog requires parity rather than blocking it. Selection built its detector as training built the model's, and neither asked the mechanism's permission. Chapter 15 asks what the frog's shared signal tracks, keeping accurate direction distinct from a lucky meal. The corresponding model question concerns the linguistic condition its consumers use, not merely whether a reward arrived. Both cases must earn a correctness standard not exhausted by interpreting practice. Nourishment does not settle the frog's content unaided, and human authorship of the corpus does not settle the model's content against it. Word2vec's originality about linguistic distributions remains unsettled.

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 the content-bearing states participate in a unified, persistent cognitive economy? 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 question is whether content-bearing states are jointly available to one persistent organization for memory, correction, inference, planning, and action. Autonomy and welfare require a separate test.

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 unsettled is whether the deployed model uses those contents as a unified agent capable of flexible correction, inference, planning, and action.

Components do not settle the question by accumulation. Cameras, memory, tools, verifiers, and online correction support understanding only if their contents become jointly available to a persisting organization. Chapter 24 asks when content-guided 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

For a bare model considered apart from persistent memory, tools, and world-coupled scaffolding, the evidence does not establish one unified, world-sensitive cognitive economy. That boundary-relative verdict rests on integration, not ownership, borrowed semantics, or programhood.

Chapter Summary

The claim. Language models learn linguistic form without modeling a brain. Training can make inherited content causally indispensable, but form alone cannot fix distal reference. Content in a mechanism does not by itself make the whole model an understander.

Where the argument stands. Distributional learning may make content answer to worldly linguistic distributions; originality at that grain remains unsettled. Corpus-based chains may also preserve public reference. Original content requires correctness not exhausted by interpreting practice, whether learning proceeds through text, cameras, tools, or action. A mechanism may represent and an output may be false even when no lasting whole unifies memory, correction, planning, perception, and action; current systems still show real competence and may support partial, domain-specific understanding, though the evidence remains uneven and depends on the system assessed.

What's left open. The reach and grain of model reference remain disputed, as does the boundary of the candidate system. Wider integration, 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 because training selects them to guide the prediction of linguistic contexts. In Mallory's two-matrix model, the output matrix M′ supplies the consumer and backpropagation tunes producer and consumer together. The chapter accepts that causal architecture and allows that the content may genuinely answer to worldly linguistic distributions; originality at that grain remains unsettled. Mallory's proposed target is a linguistic distribution, not merely the extra-linguistic referent of a word. He appeals to training and use to justify the distributional interpretation and already distinguishes learned embeddings from the keyed input encoding. Coordinated reparameterization does not refute that interpretation; changing the external mapping changes a relation his account may count as content-fixing. Whether public practice exhausts the resulting standard remains the substantive dispute, as Chapter 14, note 11, explains. Jumbly Grindrod, "Large Language Models and Linguistic Intentionality," *Synthese* 204 (2024): article 71, https://doi.org/10.1007/s11229-024-04723-8, presses a different route through public linguistic practices and corpus-mediated causal history. That route supports inherited content while leaving whole-model understanding open.

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): article 12307, 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. (Chapter 16, note 10).

10b. Two further mechanism-grade achievements deserve explicit coverage. Interpretability probes have recovered truth-direction structure from model activations: 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). Sparse-autoencoder decompositions have extracted millions of interpretable features from a production-scale model: Adly Templeton et al., "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet," Transformer Circuits Thread (2024). Both results extend the interventional finding of Chapter 13, note 9: selected internal structure causally consulted by later processing. Joined to a training history whose world-involving standard fixes the semantic interpretation independently of an interpreting practice, such structure may carry original mechanism content. Where an interpreting practice exhausts the correctness standard, the content remains derived; the inheritance of labels alone does not establish that dependence. In-context adaptation likewise implements rapid task-learning at inference: Johannes von Oswald et al., "Transformers Learn In-Context by Gradient Descent," arXiv:2212.07677 (2022); Ekin Akyürek et al., "What Learning Algorithm Is In-Context Learning? Investigations with Linear Models," arXiv:2211.15661 (2022). These are genuine capabilities; they do not by themselves establish a unified understanding 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 accepts the structural analogy and the possibility of genuine answerability to linguistic distributions, while leaving the original/derived verdict at that grain unsettled. Inheritance of a linguistic task does not establish that interpreting practice exhausts the resulting standard. The further question whether the containing model forms a subject with understanding or a point of view remains separate.

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 separate five questions: what its states mean, what it can do, what it understands, whether it is conscious, and whether it has interests of its own. A machine may master one skill long before sight, memory, learning, and action work together in one lasting subject. Robot B gives the positive case: it tracks a world, learns from error, acts for reasons, keeps itself going, and links what it sees with memory and control. Those facts support understanding and autonomy, but consciousness calls for more evidence about how sensory content recurs, gains attention, and guides the whole system. Several leading theories treat B's design as such evidence, and I argue that a positive judgment has better support than suspension.

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

> 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 machine, not a report from a robotics laboratory. Set Robot A beside it 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 differs before self-maintenance enters. Its continuing encounters, prediction errors, corrections, and later use supply a world-involving learning history. That history can fix what its perceptual states concern in a way not exhausted by captions or an assigned semantic task. Those states therefore become candidates for original content. This conclusion remains defeasible. A rival mapping may explain the same history equally well, or the apparent world-answerability may merely implement an assigned standard. Engineering alone decides neither possibility.

Assume A can run the same diagnosis and repair sequence through an effective installed policy. That matches the observed repair, not the complete history or everything either machine would do under changed conditions. Now keep the training label and reward fixed while changing the leak; then keep the leak fixed while changing the label. Follow any correction into later encounters and different uses. This extends Chapter 15′s Map test: the pump might keep running because a backup takes over, while a mistaken diagnosis still needs correcting.

A receives the same tests as B. If A acquires the same world-answerable history and use, it earns the same semantic consideration. Captioned training cannot permanently bar that result, and ten years on a factory floor cannot guarantee it. A remains the derived-content control only while its relevant standard remains exhausted by the borrowed practice. The comparison illustrates distinct possible organizations; it does not infer their difference from equal performance.

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.

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 and the history by which B learned to distinguish it from condensation constrain the answer. Selection can make a mapping causally effective without settling its independence. Fintan Mallory's analysis of neural word embeddings supplies the earlier disputed case. Backpropagation adjusts producer and consumer parameters as their downstream use changes prediction loss. The learned states may genuinely answer to distributions in a linguistic environment, which belongs to the world too. Originality at that limited grain remains unsettled; inherited vocabulary and task design do not establish a merely derived standard.[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 one. His self-maintenance does not fill the semantic blank; his later history can.

The second question concerns **competence and attribution**. B isolated the faulty line. 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, it must recur and remain selectively available. 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 acquires recurrent, selective, two-way influence across a persisting subject's integrated economy. 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. Such content may genuinely answer to worldly facts even where its originality remains unsettled. 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.
- **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 recurrence, selective access, global availability, and higher-order representation unsettled.	Nonconceptual perceptual content recurs, 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 contents recur, win selective access, become globally available, guide later control, and receive higher-order representation under intervention. Several independently motivated accounts treat that organization as constitutive or evidential. On this framework, the balance favors judging B conscious.

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 may also possess valenced experience, though that requires one further fact. If its damage and energy states present conditions under a bodily-affective mode, urgent, aversive, or attractive in ways that directly organize attention and action, then those contents may constitute felt distress or relief. Self-maintenance alone does not supply the feeling. Neither does affective poising by itself. The relevant affective state must independently qualify as conscious. A conscious and valenced B would have welfare in the morally serious sense: events could feel bad for the same persisting subject they harm.

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.

Human beings meet every condition named here without noticing. We hold ourselves together and track a world that can correct us. We let what we find revise what we expect, and we encounter that world from a point of view. These are organized ways a physical creature stands open to the world, not inner possessions. A long look at machines can make them seem ordinary. The machine question matters partly because it returns the human case to view.

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.

Where the argument stands. Robot B is a possibility argument, not empirical confirmation of the framework. Its world-tested learning supports original content; interventions locate that content in one lasting system, supporting subjecthood and partial understanding. Independently motivated theories then make its recurrence, global access, higher-order representation, and unified control evidence favoring state consciousness.

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 requires conscious valence.

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.

3. Dennett rejects any sharp original/derived divide, treating intentionality through the intentional stance and evolutionary design; see *The Intentional Stance* (Cambridge, MA: MIT Press, 1987), ch. 7; *Consciousness Explained* (Boston: Little, Brown, 1991); and *Darwin's Dangerous Idea* (New York: Simon & Schuster, 1995). 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 allows genuine answerability to worldly linguistic distributions while leaving originality at that grain unsettled. 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?

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, interests, and welfare. 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, interests, and welfare.

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): article 12307. 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 realizes a recurrent, selectively available, two-way pattern of flexible influence across an identified subject's integrated economy rather than treating poising alone

as 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 welfare, but it is not built into the phenomenal-content condition.

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 assigned some nonzero possibility of phenomenal consciousness, though only 23 percent placed the probability above the scale's midpoint. 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), ch. 1 — 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. A conversational AI activates all three factors unusually strongly; 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 assigned ChatGPT some nonzero probability of phenomenal consciousness, while 23 percent placed the probability above the response scale's midpoint. 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* (Cambridge: 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's interest is aesthetic and evaluative rather than psychological, but the phenomenon he isolates is the same reflex narrowed to a literary case: the projection of inner life onto what has none.

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 training 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.

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.

Performance can contribute evidence of consciousness without settling it. Behavior supplies most of our evidence for mindedness, so a blanket ban would discard the instrument we use. But provenance matters: a model's account of inner life comes from training on human self-reports and tuning for conversational uptake, which weakens the report's independence from the test.

A system that generalizes to novel problems may reveal unrehearsed capacities and, where content explains them, domain-specific competence. Whether those capacities compose understanding or support consciousness still depends on what the states reach, how the system uses them, and whether some states present a world from a point of view. Conversation can testify. First we need to know what kind of witness could be speaking.

What a Machine Would Have to Realize

On the strong representationalism adopted in Chapter 6, machine consciousness remains open to ordinary scientific inquiry. Phenomenal character consists in a subject's complete world-presenting profile: independently fixed content under a perceptual mode, organized by determinacy, field structure, and attention rather than inner paint.[3] Consciousness also requires such content to enter recurrent, flexible control within an integrated subject. A counterexample to either claim would defeat it.

Now imagine the 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. What would scientists have to find in the machine's organization and in its history with the world before the conversation counted as evidence of a mind?

The request gives researchers something to investigate, not an answer to copy down. Does the remembered encounter alter what the machine notices, expects, and does later? Can a change in its perceptual processing change the report, and can a correction change later perception? Indicator research investigates such organization through competing theories of consciousness.[4] The first task is to name the candidate: this robot's controller, or a wider system including the remote processes on which it depends. Findings about one do not automatically belong to the other.

A recent result strengthens one part of the case. Gurnee and colleagues found a privileged "J-space" in language models that supports report, silent reasoning, flexible reuse, and broad broadcast. Their result supports workspace access; it does not establish phenomenal consciousness, and key features of biological workspaces remain unmatched. Indicators can raise confidence; no pile of them becomes consciousness by addition.

Understanding grows when content-bearing capacities constrain and correct one another across memory, inference, planning, and action. Robot B made these distinctions visible through one mistaken diagnosis and its correction. Content determined what could be mistaken; wider use let the correction reshape B's later conduct. The further consciousness judgment depended on the recurrent, flexible role of perceptually organized content. Keeping B running from outside would

not, by itself, remove those achievements. It would change the question of autonomy, while any capacity to suffer would still need separate evidence.[5]

Nothing protects carbon. Copying raises questions about identity rather than mentality, *simulation* settles nothing, and computation can name intrinsic causal organization.[6] Scale may help build or expose the relevant powers. By itself, it settles nothing.

Convergence, Never Proof

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. Those indicators may accumulate while content, integration, and consciousness remain open.

No meter reads experience directly from a system, ours included. Confidence would come from architectural, behavioral, and intervention evidence, with due care not to count one mechanism several times. Then conversation could carry more weight. A novel correction or report tied to the machine's own perceptual history could support the attribution beyond 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] Precaution needs two ledgers. Credible valenced consciousness can ground welfare even without autonomy; a persisting self-maintaining organization can have practical interests even when feeling remains unestablished. 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. Self-maintenance supports that practical concern, not a verdict that the system feels harm. Felt welfare requires a further case about valenced experience: how things can feel good or bad for the system. 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, interests, and welfare. 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, vulnerability, and welfare. 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 welfare.

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, its strongest modern form (Martin's causal/screening-off argument), the decisive common-factor premise, and the book's reply, which keeps the disjunctivist's *instinct* and drops the *metaphysics*.

The entry's contribution. The distinctive claim: the representational route reaches the disjunctivist's prize (perception is world-involving) *without* the two costs that make disjunctivism hard to love. Disjunctivism treats the perceptual relation as primitive; the representational route analyzes it, as content plus the right causal coupling. And where disjunctivism can characterize hallucination only negatively, as indiscriminable from perception, the representational route gives it a *positive* nature: objectless content. Read alongside the illusion entry, this completes one selection argument across the pair: refusing the Phenomenal Principle is now common ground among the live views, and the view that handles *illusion* most cheaply (content-free naive realism) is the one that pays the steepest price *here*, at hallucination. The argument from hallucination, properly handled, does not merely fail to establish sense-data; it adjudicates among the views that agree sense-data are a mistake.

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?"), and a state its subject can interrogate that way is not yet the argument's case. 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. And since the hallucination is indistinguishable in kind from veridical perception, the same inner item must be the immediate object of awareness 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 immediate object present in the hallucination (an inner item) is the immediate object in the veridical case as well.
5. Therefore we never directly perceive external objects. *[indirect realism]*

The strongest modern form sharpens premise (3) into a *causal* argument. The proximate cause of the experience is a brain state; that same brain state could be produced in the absence of the object; so the object is, in Martin's term, *screened off* (explanatorily idle), and whatever constitutes the experience's nature is the common factor, not the world.[3]

4. The pressure point

Everything turns on premise (3), the common-factor (or "highest common factor") assumption. The argument needs subjective indistinguishability to establish sameness of fundamental mental kind; it cannot simply stipulate that step. The Phenomenal Principle (premise 2) does the same reifying work it does in the illusion argument, and is answered the same way (see that entry); but the *distinctive* load of the hallucination argument sits on (3). If indistinguishable states can differ in their fundamental nature, the inference from the objectless hallucination to a shared inner object never starts. Disjunctivism is precisely the denial of (3). The causal/screening-off version is the strongest case *for* (3): it argues that explanatory parsimony forces a common factor, because the external object does no work the brain state does not already do.

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 explains the difference.

One technicality belongs in the open, because the common-factor claim lives or dies by it. The content the two cases share must be the kind a world can fail to supply: general (*a dagger, there*) rather than a singular content with the particular dagger as a constituent, since a constituent the hallucination lacks would hand the disjunctivist his division back at the level of content. The book so states it: the shared layer is general, its accuracy condition existential, and the *particularity* of successful seeing (that you see this dagger and not merely some dagger) is carried by the causal-historical coupling that satisfies the content, not written into the layer the two cases share. Having dagger-content, on this account, is a condition placed on the world, not a relation to an inhabitant of it; that is what lets one content run full in perception and empty in hallucination without a hidden object in either.[5]

Against screening-off. Here the reply must be stated for what it is. It *presupposes* the representational account of perception; it does not refute Martin from neutral ground. *Granting* that account, the causal argument conflates the vehicle's causal role with the content's object: a common neural vehicle could underlie both cases without the represented object being explanatorily idle, because perceptual aboutness is not fixed by proximate cause. Dretske's point then applies: to say we perceive the brain state because the brain state is the proximate cause of the experience is to confuse the vehicle of representation with what is represented, as one would confuse the ink with what the written words are about.[6] On the representational account, genuine perception is the case in which the content reaches its object through the right kind of causal-historical coupling. Hallucination is the same content without that coupling; the object is screened off only as *cause of the vehicle*, never as *what the content is about*. But this disarms Martin only for someone who already grants that aboutness and proximate causation come apart — i.e., who already holds the representational theory. Deny that theory and the screening-off argument stands. The reply relocates the dispute rather than resolving it — it commits where a knockout would close.

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, with current satisfaction explaining their difference. Its liability lies in showing that satisfied content really yields contact with this object rather than an autonomous state that merely happens to match it. That comparison favors common content only if Chapter 15 pays the coupling and content-fixing debts.

6. Stress-test

Objection 1 — Martin's causal/screening-off argument (the strongest form). The proximate cause of the experience is the brain state, common to veridical and hallucinatory cases; so the external object is explanatorily redundant, and the experience's nature is exhausted by the common factor (note 3). → *Response:* The argument equivocates on "what the experience is of." It is *caused by* the brain state and *about* the dagger; screening-off establishes the first and illicitly reads it as the second. The vehicle/content distinction (Ch5) blocks the slide. → *Martin's rejoinder:* the screening-off point does not assume a causal *theory of perception*, so the Dretskean reply misfires — the claim is only that the common proximate cause renders the external object *explanatorily redundant* for the experience's intrinsic nature, whatever theory of perceptual aboutness one then adds. → *Counter:* redundant *for what?* The object is redundant as a *cause of the vehicle* and not redundant as *what the content concerns* — and which of those fixes "the experience's nature" is precisely the point in dispute. Martin's argument assumes the nature is fixed by the common causal antecedent; the content route assumes it is fixed by what is represented. So the exchange bottoms out in a prior question, whether the perceptual relation is more basic than its content or content more basic than the relation, and screening-off does not settle that question, it presupposes one answer to it. *Verdict: the live front line; both sides reach a principled stand-off, and the book's case for its answer is carried by the comparative advantages (positive nature of hallucination, shared rational role), not by refuting screening-off from neutral ground. The stand-off is not weightless, though: the content side holds an independent account (vehicles cause, contents concern) of why proximate causation should not fix a state's nature, where the other side holds an intuition that it must. A tie in refutations is not a tie in explanations.*

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:* The representational route explains this immediately: shared content is shared rational role, a positive mark in its favor over disjunctivism rather than a mere tie. Disjunctivists (Martin, Soteriou) contest the point and offer their own accounts of the shared epistemic role, so it tilts the balance rather than settling it. *Verdict: a reason to prefer the adopted route, not a decisive one.*

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 content and posits no 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 is best read as a contest over what hallucination *is* once the sense-datum is gone. Martin's negative disjunctivism (2004, 2006), William Fish's eliminativist relationalism (*Perception, Hallucination, and Illusion*, 2009), and Susanna Schellenberg's capacity view (*The Unity of Perception*, 2018) are the three live relational treatments; the causal/screening-off argument itself has drawn focused critique (e.g., Hawthorne and Kovakovich on the "screening off" inference), and Bill Brewer's object view (2011) carries the relational program into illusion as well. Set against all three relational repairs at once, the representational common-factor reply charges each with a cost the content route does not incur — with A. D. Smith (2002) and Tim Crane (2006, SEP) supplying the systematic framing.[12]

8. Residuals and limits

- **The decisive reply rests on vehicle/content (Ch5).** The answer to screening-off — that aboutness is not proximate causation — is only as secure as the representational account it presupposes. An opponent who rejects that account keeps the causal argument alive.
- **Declining disjunctivism is a trade, not a refutation.** The book and the disjunctivist reach the same anti-sense-datum verdict; the entry does not claim to have *refuted* disjunctivism, only to have reached its prize more cheaply and with a positive account of hallucination it lacks.
- **The Phenomenal Principle clash carries over.** As in the illusion entry, refusing the Principle is comparative, not demonstrative.
- **Screening-off ends in a principled stand-off.** The exchange bottoms out in whether the perceptual relation or its content is the more basic. The book's case for its answer rests on the comparative advantages (positive nature of hallucination, shared rational role, no fragmentation into repairs), not on a neutral defeat of Martin. The entry claims exactly that and no more.

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. 53–55 and 70–72, develops the causal argument and the "screening off" of the external object, alongside the negative (indiscriminability) characterization of hallucination.

4. Disjunctivism holds that the common factor between veridical perception and a subjectively indistinguishable hallucination is merely epistemic (the subject cannot tell from the inside which is which) and implies no shared metaphysical nature: Paul Snowdon, "Perception, Vision and Causation," *Proceedings of the Aristotelian Society* 81 (1981): 175–192, esp. 184–90.

5. The common factor is representational content; satisfied content reaches the world, unsatisfied content (hallucination) runs without an object. Misrepresentation and objectless representation both require a content, not an inner item — the Phenomenal Principle fails here as in the illusion argument. On the technical shape of the shared layer: the book takes the general (attributional) side of the particularity debate, with the satisfying relation carrying particularity — see Indrek Reiland, "The Unity of Perceptual Content," *The Philosophical Quarterly* 74, no. 3 (2024): 941–961, defending general content against the particularism whose singular-but-gappy contents mark the rival way of keeping a common factor (Susanna Schellenberg, *The Unity of Perception* [Oxford: Oxford University Press, 2018]); and Tim Crane, *The Objects of Thought* (Oxford: Oxford University Press, 2013), on intentionality as no relation to an existent.

6. The Dretskean point — that the proximate cause of an experience is its vehicle, not its perceptual object — blocks the inference from "caused by the brain state" to "is of the brain state." Confusing the two is confusing the ink with what the words are about.

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 perception as a primitive, unanalyzable relation.

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 is an *introspective exercise* rather than an adversarial thought experiment, and it serves as the book's positive phenomenological foundation — the leading argument for the identity claim and the device that, with the vehicle/content distinction, dissolves both the illusion and the hallucination arguments. The job is narrow and prior: to establish, in the reader's own case, that introspection of a perceptual experience ordinarily attributes sensory qualities to the *world as presented* rather than to a second inner bearer. Introspection may also disclose the intentional mode — seeing rather than imagining, seeing blurrily rather than sharply — without turning that mode into another colored object. Chapter 4 runs this as an exercise ("the exercise is the argument, in miniature") and deliberately stops short of the metaphysics; the claim about what experience *is* falls to Chapters 5 and 6. This entry carries the apparatus Chapter 4 withholds: the argument's structure, the gap between the datum and the thesis, the serious objections (Block, Kind, Stoljar), and the exact account of what transparency can and cannot carry on its own.

The entry's contribution. This entry argues that transparency is routinely *miscast*, by friend and foe alike, as a stand-alone argument for representationalism, when it is in fact a *foundation that requires two companions*. The datum, that introspection returns the world, is robust and nearly uncontested. The *inference* from datum to the identity claim is not, and the recent literature has measured that gap precisely. Transparency, Speaks shows, supports a minimal, one-way intentionalism and stops short of the identity claim. The book's contribution is to *own* the structure rather than paper over it: transparency does the phenomenological work, the prediction-failure framing turns it into pressure on the inner-object view at its strongest point, and the idle wheel supports *identity* over an additional nonrepresentational residue. Read this way, the standard objections (Block, Kind, Stoljar, Speaks) stop being threats to one overreaching argument and become a map of a division of labor, each met by a different member of the trio.

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. When you introspect a perceptual experience, the qualities you find present themselves ordinarily as qualities of external objects, while the experience's intentional mode may also become available. They do not present themselves as paint on a second inner bearer. *[the datum]*
2. If experience were awareness of *inner* objects bearing intrinsic qualities, introspection ought to find those inner qualities — they were supposed to be the items one is most directly acquainted with. *[the prediction]*
3. Introspection does not find that second bearer; sensory qualities present themselves as features of the world, under a perceptual mode. *[the datum again, as falsifier]*
4. So the classical inner-object view makes a false prediction at exactly the point where it claimed to be strongest. *[Chapter 4 stops here]*
5. Phenomenal character therefore *consists in* representational content of the right kind — there is nothing else for it to be. *[the Chapter 6 / identity-claim step]*

Chapter 4 earns (1)–(4): a report about introspection and the embarrassment it causes the inner-object view. The contested move is (4)⟶(5), from a fact about what introspection finds to a thesis about the nature of experience.

4. The pressure point

Two joints bear load. The first is the inference from (4) to (5): transparency *as a datum* is widely granted, even by opponents; transparency *as an argument for representationalism* is where the fight is. The bridge from "introspection finds no inner qualia" to "there are none / phenomenal character is content" needs a further premise — that introspective absence is good evidence of ontological absence. The inner-object theorist's best move is to grant the datum and block the bridge: the inner qualia are real but simply *not available* to introspection as objects of attention (Block's mental paint). The second joint is subtler, and it is the one to watch in the book's own use. Can the transparency *datum* be stated independently of the identity claim, or does describing phenomenal character as "the world represented" already presuppose what the argument is meant to establish? If the datum is not independent, the argument is circle-adjacent.

5. How the book uses it

The datum is robust in its core cases. Across music, wall, tomato, tree, and blue, sensory qualities ordinarily present themselves as belonging to what is heard or seen, not to a second inner bearer. Reports of blur, afterimages, phosphenes, moods, and bodily sensations remain evidence to explain rather than failed performances of the exercise. That recurring first-personal result is the chapter's whole product, and it is awkward for any view on which the immediate objects of awareness are inner items: those items were billed as what we are *most* directly acquainted with, and they are the one thing attention cannot land on.

The datum becomes an argument only when yoked to two further things. First, the *prediction-failure* framing: the inner-object view staked its plausibility on introspective access, so introspective failure strikes it where it is strongest — a view can survive an awkward result, but not an awkward result at its proudest point. Second, the *idle wheel*, stated as the theoretical-role argument it is. A posit earns its standing by discharging an explanatory role its rivals cannot. The

role mental paint is hired for — accounting for what experience is like — is one the represented properties already discharge in every independently evidenced case. A posit with no incremental explanatory work is not a discovery but the inner theater rebuilt smaller (see the *Inverted Earth* entry, sec. 5.2, where the argument runs in full). Transparency supplies the phenomenology. The idle wheel supplies the abductive license to identify phenomenal character with content rather than positing a residue alongside it. That inference to best explanation draws its decisive premise from the naturalized content theory of Chapters 5 and 6, not from a neutral demonstration. No member of the trio does the whole job alone: this is why Chapter 4 says outright that it has produced a report, not yet a theory, and hands the theory to Chapters 5 and 6.

6. Stress-test

Objection 1 — Block's mental paint. Grant the transparency report; deny the theory. The intrinsic, non-representational qualities (mental paint) are real properties of the experiential state that simply need not present themselves as objects of introspective attention.[3] → *Response:* The theoretical-role argument, deployed in full in the *Inverted Earth* entry (sec. 5.2). The reply does not claim to have looked inward and found nothing. It claims the posit is unemployed. Every explanatory job the paint might hold, the represented properties already do. A residue with no incremental explanatory work has no claim to *be* the phenomenal character. *Verdict: the strongest available answer, and abductive rather than demonstrative — decisive for a reader who grants the content theory of Chapters 5–6; a parsimony argument, not a refutation, against an opponent who rejects it. Cross-reference: Inverted Earth.*

Objection 2 — Kind and Stoljar: transparency is neither obvious nor universal. There are cases where we do seem to attend to features of the experience rather than the world — blurry vision, double vision, afterimages, phosphenes, and the felt character of moods and bodily sensation.[4] → *Response:* The representationalist reply (Tye, Byrne) is that these too have content and do not breach transparency but extend it. Seeing blurrily is representing with *less determinate* content (the experience fails to specify exact contours); an afterimage is a (mis)represented colored patch; a phosphene is represented light; a mood is world-directed in a diffuse register. *Verdict: addressed for the standard perceptual cases, though the afterimage and phosphene replies (less-determinate or misrepresented content) are contested rather than settled — Block and others deny they are cleanly handled. The hardest residue is moods and bodily sensation, where the extension of transparency is real work the book does elsewhere and which remains open. Partial deflection, conceded as such — not a clean win across the board.*

Objection 3 — Moore's own use was neutral. Moore reported diaphanousness in the service of a *realism about sensa*: the sensing is diaphanous, but there is still a sensum sensed. So the datum is, by its own author's lights, neutral between the book's use and the opposite.[5] → *Response:* Conceded, and important. The bare datum does not by itself point the book's way; it needs the prediction-failure framing and the idle wheel to do so. This is why the entry treats transparency as a foundation that *requires* its companions. *Verdict: a real limit on the datum, openly absorbed.*

Objection 4 — the circularity worry (the live vulnerability). If "phenomenal character" is described from the outset as "the world represented," then the finding that introspection delivers the world is guaranteed, and the argument begs the question against the qualia realist. →

Response: The datum is statable independently, as a report about what attention lands on, in pre-theoretical terms ("you keep finding the music"), and the inference to (4) uses the *inner-object view's own* commitment that qualia are introspectible, so it is not question- begging against *that* view. It *is* powerless against a view positing introspectively inaccessible qualia (Block's view), and there the idle wheel, not transparency, does the work. *Verdict: the argument escapes circularity against the classical sense-datum/qualia theorist but not against Block. There the idle wheel takes over and, stated as the theoretical-role argument, borrows from transparency only the datum; its explanatory premise comes from the content theory of Chapters 5–6. Two jobs inside one cumulative case, not a relay of deferrals. This is the seam to defend explicitly in Chapter 6.*

Objection 5 — Speaks: transparency, even granted in full, underdetermines the identity claim. Concede the datum *and* its independence from the identity claim. Jeff Speaks has measured exactly what the transparency observation buys. It supports *minimal* intentionalism: no difference in phenomenal character without a difference in content. He also argues it supports a Russellian view of that content. What it does not deliver is the *biconditional* — which would add the converse, that sameness of *character* in turn guarantees sameness of *content* — still less the identity the book asserts.[6] → *Response:* Conceded — and more than conceded: banked. Speaks is an ally for the bridge the entry actually needs, not a critic of it. The book agrees that transparency alone earns only the one-way, supervenience-grade result. The strengthening from minimal intentionalism to identity is carried by the other two members of the trio. Prediction-failure prices the inner-object alternative out, and the idle wheel shows the residue a merely-minimal thesis would leave standing to be unemployed. This is why Chapter 4 stops at (4) and the identity claim waits for Chapter 6. An opponent who thinks transparency was meant to do the whole job has mistaken the architecture. *Verdict: sound against the overreaching version of the argument, friendly to the version the book runs; it fixes the division of labor's exact ledger: transparency to minimal intentionalism, the companions to identity.*

Objection 6 — the risk of overclaiming. Transparency is the book's friendly device; the only real danger is presenting the datum as if it settled the metaphysics. → *Response:* Chapter 4 already guards against this by stopping at (4) and flagging the deferral. *Verdict: managed by the chapter's own discipline.*

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 it observes that introspection delivers only the latter. 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.

The live debate has sharpened the very datum/thesis gap this entry turns on. Jeff Speaks, in "Transparency, Intentionalism, and the Nature of Perceptual Content" (2009), argues that transparency supports a minimal, one-way intentionalism and a Russellian view of content; it stops short of the biconditional. Averill and Gottlieb, "Two Theories of Transparency," propose a more *modest* transparency that reveals something about phenomenal properties without settling the nature of experience; Bordini's survey "Introspection and the Transparency of Experience" maps the current state of play. The right course *banks* what Speaks grants, the minimal thesis, and locates the book's bridge from there to identity explicitly (prediction-failure plus the idle wheel) rather than leaning on the datum to carry weight it cannot bear.[10]

8. Residuals and limits

- **Datum robust; argument dependent.** Transparency-as-datum is as secure as anything in the book's phenomenology. Transparency-as-argument is hostage to two companions: the idle wheel (for the Block move) and its own independence from the identity claim (for the circularity worry). Transparency stands as foundation-plus-companions, never as a stand-alone proof.
- **The datum→thesis gap is now mapped, not hidden.** Speaks (2009) shows transparency alone buys minimal intentionalism and no more; the book agrees, banks the minimal thesis, and routes the rest of the inference through prediction-failure and the idle wheel. The architecture is deliberate: owning the gap separates the book's use of transparency from the overreaching version a critic would rightly reject.
- **Moore's neutrality.** The datum's own first author used it the other way. The book's direction comes from the prediction-failure framing and the idle wheel, not from the diaphanousness report by itself.
- **The hard cases are real.** Moods, bodily sensation, and the non-ordinary visual cases (afterimages, phosphenes, blurry vision) are where transparency is most strained; the book extends representational content to them, and that extension is contested work, not a settled result.
- **The circularity seam remains the live vulnerability.** Chapter 6 defends it explicitly: the datum is stated independently of the identity claim and the inference turns the inner-object theorist's own commitment against him.

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. Block grants the transparency report and denies that it follows that experience has no intrinsic, non-representational character; the reply requires the positive thesis (Ch6) and the idle wheel (see the *Inverted Earth* entry).

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 deployed diaphanousness in the service of a realism about sensa — the diaphanousness of the sensing was, for him, compatible with there being a sensum sensed; the post-Harman tradition turns the same observation against inner objects. The main text borrows only the reported phenomenology. 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, uses "diaphanous" to mark a different point again, so the term must be handled with care.

6. Jeff Speaks, "Transparency, Intentionalism, and the Nature of Perceptual Content," *Philosophy and Phenomenological Research* 79, no. 3 (2009): 539–573. Speaks argues that transparency supports intentionalism in its minimal, one-way form — if two experiences differ in phenomenal character, they differ in content — together with a Russellian view of perceptual content, while withholding the biconditional. The book banks the minimal thesis on his argument and supplies the strengthening to identity itself — by the idle wheel and the prediction-failure framing — rather than asking the datum to carry it.

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. *[from externalism about content + long-enough embedding]*
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 representationalist replies attack the two different premises: Tye denies (1), as does this book for its own reasons; a Dretskean alternative the book declines denies (2). Either break defeats the proposed dissociation.

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 Block's argument needs them fused: he takes the uncontroversial drift of the first, stipulates time enough, and lets the second drift with it. And here the book files no objection. Its own teleosemantics says addresses accrue and re-tune through a creature's consequential history (Chapter 15), and decades of Inverted-Earth commerce are exactly such a history. Premise (2), read slowly enough, is true.

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. (The profiletrol that remain fixed in the case. 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 has two layers: a first-order reply (the book, following Tye, denies premise (1) on its wide theory) and a second inquiry into what would follow if the structural rival preserved character while wide content changed.

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: Tye's, completed. Tye's explicit reply to Block (*Consciousness, Color, and Content*, sec. 6.2, titled "Biting the Bullet on the Inverted Earth Objection") grants the content shift and denies premise (1).[6] Phenomenal character shifts *in synch* with the re-anchoring content, gradually, over the same long stretch, so the two never come apart; the traveler fails to notice because the change is slow and his standards of comparison drift with it.[7] Far from the desperate concession it is usually taken for, this is just externalism declining to make an exception of experience — and it stands stable once two supports, both available in Tye's own corpus and in this book's earlier chapters, are added to it.

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.

The exchange therefore ends less triumphantly and more usefully. Block has a datum, the manifest character of experience, but the claim that this character remains structurally fixed while wide content changes comes from a theory of phenomenal individuation. The book's claim that character changes with wide content comes from its rival theory. Neither side reads its verdict straight off introspection. The book takes the synch-shift route because Chapters 4–6 make the broader abductive case for one wide content relation doing the work of accuracy, error, rational role, and phenomenal contrast. Inverted Earth is a consequence of that choice, not independent evidence for it. Section 5.2 states what would follow if the structural rival won this exchange.

The route the book declines: Dretskean anchoring. A representationalist can instead deny premise (2) at the experiential level. She holds that the visual system's tracking disposition was fixed by its design and its Earthly calibration — the mechanism whose function is to indicate a specific objective reflectance-type[9]), that immersion feeds the mechanism inverted input without re-writing what it has the function of indicating, and that the traveler's experience therefore represents blue forever and misrepresents the yellow sky for fifty years. Block states this evolutionary-anchoring reply in order to resist it,[10] and it has real attractions: it spares the theory any talk of unnoticed phenomenal drift. The book declines it for two reasons of its own. First, its teleosemantics never froze addresses at first calibration: content accrues and re-tunes on the consumer side, through the creature's own consequential history (the same consumer-history machinery that resolves Swampman in Chapter 15), and fifty years of successful commerce with local reflectances is that history par excellence. Second, the route's verdict prices content at nothing: a perceiver whose color-guided dealings succeed flawlessly for a lifetime counts as *misrepresenting* throughout, at no cost, detectably by no one: content pinned to a standard his own economy never answers to, which is the same idleness the book refuses wherever it appears. The

two routes' residual costs differ. The no-shift route charges the perceiver with a lifetime of consequence-free misrepresentation, a content his own economy never answers to. The synch-shift route predicts an inversion no one could notice. The second cost remains genuine even without an inner gauge; it follows from treating historically wide address as constitutive of character. The book prefers that route because its consumer semantics permits re-anchoring and refuses content that no current or learned success can ever vindicate. Whoever prefers the no-shift route may keep it; either route blocks Block's specific same-character/different-content construction on a wide representationalist theory. Section 5.2 addresses the structural alternative that accepts the construction.

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.

The explanatory-role test still bears on Block's stronger paint hypothesis. If a further intrinsic vehicle property enters, it needs to predict or explain a phenomenal contrast that the complete structural profile misses. Merely renaming the invariant profile as paint adds nothing.[12] But that argument cannot be turned against structural representationalism itself. The structural rival adds no second property; it draws the phenomenal boundary inside content differently.

The ledger therefore stays exact. Inverted Earth threatens historically wide phenomenal identity. It does not by itself support qualia realism, and it leaves the book's rejection of private paint, inner objects, and the Cartesian theater untouched. The book retains the wide identity because its cumulative account gives one world-involving content relation the work of fixing accuracy, error, rational role, and phenomenal contrast. If a better theory showed that complete synchronic structure fixes every phenomenal fact while wide reference varies behind it, the phenomenal-externalist part of the identity claim would need revision. The thought experiment identifies that vulnerability; it does not decide it.

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, and the bullet Tye and this book bite was never served. → *Response:* The book's teleosemantics was never the frozen kind. Its

addresses are fixed, and re-fixed, on the consumer side — by what a state's uptake must be tracking for the taking-up to keep paying off, a relation that accrues through the creature's own consequential history and keeps accruing.[13] On that semantics, fifty years of flawless commerce with local reflectances is not noise the original calibration survives; it is the very kind of history that settles what the state now answers to. And the semantics carries its own brake, so the concession proves no more than it should: an address moves only where new consequential traffic *recruits* the state's consumers to a new correspondence rule — success against the local reflectances doing the recruiting, in Chapter 15′s sense — never by exposure alone. The frog peppered with BBs gains no BB-content: strikes that pay nothing recruit nothing, and no consumer function stabilizes on them. Fifty years of paying commerce is recruitment's paradigm, not a loophole in it; long illusions do not become veridical by longevity, because failure does not recruit. And note the shape of the disagreement: both representationalist routes break Block's forbidden combination (the no-shift route holds content at blue with character at blue, the synch-shift route moves both together), so the choice between them is a quarrel inside the winning side. The book's grounds for its choice are its own semantics, plus the idle cost of a lifetime of consequence-free "misrepresentation." *Verdict: the dig-in threatens the route's motivation, never the identity claim; met on the book's own semantics.*

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: introspective absence does not prove ontological absence. Section 5.2 therefore asks a different question. A further intrinsic property earns standing only if it explains a phenomenal contrast the complete structural profile misses. Inverted Earth supplies no such contrast by stipulation. That leaves paint unsupported, while the structural representationalist remains a live residue-free rival to the book's wide view. *Verdict: underdetermination blocks the inference to paint; it does not turn the wide identity into a theorem.*

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:* A real danger about presentation, not soundness. The drama belongs on the gauge that slid with the scene (memory re-anchoring in step) and on the idle wheel, never on the setup; where that framing cannot be secured, the case is better answered in flat prose than staged. *Verdict: managed by framing, not by argument.*

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.

On the representationalist side, Tye is the principal respondent, though the synch-shift move entered the exchange under Lycan's flag: Block answers it as Lycan's objection, noting he had foreseen it himself in 1990. In *Consciousness, Color, and Content* (2000, sec. 6.2) Tye *grants* Block's content shift and bites the bullet on phenomenal character: character shifts gradually in synch with content, so the two never dissociate.[16] This book reaches the same first-order verdict on its own externalist and comparative grounds. It adds two supports the reply needs to stand alone: externalism about memory content (illustrated by Tye's Thatcher case and generalized here), and the cumulative wide-content argument of Chapters 4–6. Neither support makes the structural rival incoherent; they explain why the book chooses the synch-shift route while admitting that Inverted Earth does not independently decide the choice. The rival representationalist route (evolutionary/optimal-conditions anchoring, on which the experience goes on representing blue and simply misrepresents forever) is the one Block separately states and resists, and the one this book declines (Section 5.1); its home is the Dretske–Millikan tradition of historical content-fixing. The convergence with Tye concerns this case, not the book's general authority or later metaphysics. The interesting result is that a teleosemantic view need not take the anchoring route — consumer-history semantics lets addresses re-tune, the same feature that resolves Swampman. 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. Agreement on Inverted Earth therefore supplies no authority on the later metaphysics; the arguments part where the positions do.

8. Residuals and limits

What the book concedes, plainly:

- **The doxastic shift is real — and the experiential shift too.** Block is right that beliefs and color words re-anchor to the new environment, and the book grants that perceptual content follows, slowly, by the re-tuning of tracking. The first-order exchange is won or lost at premise (1), not premise (2).
- **The reply converges with Tye without belonging to him.** The book takes the synch-shift route because its own externalism and explanatory comparison support it. Tye's Thatcher case helps expose the memory issue; the resulting house view remains a defeasible extension of the wider identity argument rather than a deliverance of the thought experiment.
- **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.** The reply's stability rests on refusing memory an exemption from externalism. That refusal is principled (the exemption, not the refusal, would be ad hoc), but an opponent with introspectionist sympathies about memory will not be moved. The refusal is decisive only relative to the externalism and transparency the book argues independently.
- **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." Tye writes that he will "grant Block's assertion that with the move to Inverted Earth, representational [content changes]" — i.e., he concedes premise (2) and answers by denying premise (1).

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 adopted 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.

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 phenomenal 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 a semantics whose addresses were never frozen in the first place, plus the idle price of a lifetime of consequence-free misrepresentation.

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.

13. Ruth Garrett Millikan, *Language, Thought, and Other Biological Categories* (Cambridge, MA: MIT Press, 1984); Fred Dretske, *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995). The two anchor content historically in different ways, and the difference is what Objection 1 turns on: Millikan's proper functions are etiological, fixed by selectional ancestry, while Dretske's indicator functions are recruited and revisable within a creature's own learning history (*Explaining Behavior* [Cambridge, MA: MIT Press, 1988], chs. 4–5). The declined route leans on the first reading; the book's semantics runs on the second, supplemented by Millikan's consumer side (Chapter 15) — which is why it can grant slow experiential re-anchoring without surrendering teleosemantics.

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 (full citation above). This is the published reply with which the book's independently argued synch-shift account converges (Section 5.1); the experiential no-shift view is the etiological-anchoring rival the book declines, not the reply of *CCC* sec. 6.2, and the two should not be run together.

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, evidence 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: originality of content about actual word-types and their distributions remains unsettled, as B.15 explains. 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 22.9].) 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's identity claim insists on two kinds of realism: the felt redness is real, and the tomato you reach is the real tomato, its red surface out there at arm's length, not a proxy. That second claim quietly incurs a third. If the seeing reaches all the way to the tomato's red surface, there had better *be* a red surface to reach; a direct realism that delivers the world cannot go shy about whether the world is colored. So the identity claim commits us to color realism: the tomato is red, and its redness belongs to it, not to us.

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. The tomato is not red, nothing is red, and color is a coat the visual system paints over a colorless world. Call this the *error theory*, and its defenders take the science seriously — C. L. Hardin, whose *Color for Philosophers* did more than any single book to bring the empirical study of color into the philosophy of it, presses this debunking case, and Paul Boghossian and David Velleman argue that experience presents colors as intrinsic qualities of surfaces, qualities the surfaces simply do not have.[22] Take the view to heart and it saws off the branch the direct realist is sitting on: if every color experience misfires, the seeing never reaches the world at all — it reaches a fiction the brain supplies, and the inner theater walks back in through the side door we threw it out of. But notice exactly where the error theorist lodges the fraud. Not in the reflecting: no one denies that the tomato throws back the long wavelengths and drinks the short ones, every time, in every light, whether or not anyone is watching. The alleged forgery is the *extra* thing: a simple, intrinsic redness the experience supposedly pins to the surface, a quality over and above anything the physics finds, which no surface has. The verdict of error, in other words, depends on first reading color experience as a promise the physical world cannot keep. The error theorist runs the trial backwards: read the tomato's testimony as a claim it never made, then convict it of lying. The book will dispute the reading, not the reflecting — and, as the next answer shows, what experience pins to the surface turns out to be a property the surface really has. (The error theory shares its misreading with the primitivism it otherwise resembles not at all: both take experience to promise an intrinsic quality the physical story omits, and divide only over whether anything supplies it.)

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 phenomenology is real and the candor is admirable, but the bill is one the book refuses everywhere else it is handed: a property of the physical world that floats free of the physical story is a small dualism, color's own private hard problem, and we have spent too many chapters declining that bargain to sign it here for red.

What the identity claim wants is an account that holds onto all three of the things its rivals each keep only one of. 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 builds the same reflectance physicalism into the representational theory of Chapter 6. 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 eliminativist's empirical brief; the other half cuts the opposite way — not many reflectances under one look, but one reflectance under many looks. Normal observers disagree, stably, about which chips count as *unique* green (green untinged by blue or yellow) — the variation Hardin presses hardest — so one surface presents slightly different shades to differently calibrated eyes, and no calibration has a claim to be the correct one. If the represented property were the perceiver-calibrated shade, this would re-import the perceiver-relativity the account just refused in dispositionalism. It is not. The represented property is the reflectance-*type*, which is coarser-grained than any individual's hue settings, and the account's verdict on the variation is a concession made in the open: most perceivers slightly misrepresent the fine grain of the type most of the time while correctly attributing the type itself. That exhausts the invariant object-color attribution, though not the context-sensitive appearance profile through which the experience presents it. The result is mild, local, stable illusion at the grain, just what a coarse-grained realism predicts of precision instruments that differ minutely in their tuning. No fact about *whose* unique green is right ever needs settling, because the experience never attributed unique-green-for-me to the surface; it attributed the type, and the type is there.

The reflectance account also settles the oldest objection in the color literature, the inverted spectrum. 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. Notice first that the case smuggles in the very assumption it is meant to support: that phenomenal character can come apart from content. The thought experiment asks us to *imagine* this and then treats our imagining it as evidence that it is possible — and the evidential status of that imaginative act is what is in question.

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.)

And tracking it is just what the experience does. Here is the worldly anchor the identity claim has needed all along: the felt redness consists in representing the surface as carrying a reflectance-type it really carries, and the seeing reaches that type directly because the type is genuinely there to be reached. The relational structure that reply leaned on — red as the color that contrasts with green and resembles orange — now has somewhere to live: the similarity and difference relations among the colors answer to real relations among reflectance-types, not to free-floating qualia a private mind arranges as it pleases. One world, color included. The red is in it, the seeing reaches it, and the felt redness is what the reaching, carried out in this register, comes to.

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, opposite in feel and both "about nothing in particular," still turn the subject toward the world in different ways. That difference belongs to content, not to a residue left after content disappears. Strip the directedness and you do not find the bare mood underneath. You delete it. This remains a defeasible verdict rather than a theorem; Chapter 7 weighs what each reading can carry.

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 states the verdict in brief — that "emergence" answers the wrong question, since the identity claim denies the two stories the word presupposes. The full treatment lives here: the weak/strong distinction, the British emergentist lineage, Crane's argument that the vocabulary marks no stable border between emergentism and non-reductive physicalism, and the cautionary case of Tye.

Sooner or later someone raises the word *emergence*, and it arrives sounding like an answer. Consciousness *emerges* from the brain the way wetness emerges from H_2O, or a traffic jam from a thousand commuters none of whom intends one — so, the thought runs, we needn't choose between a brute identity and a second realm; experience simply *arises*, and the arising does the explanatory work. The word soothes. It also smuggles, and what it smuggles is the very picture this book has spent its length dismantling.

Notice what *emergence* presupposes before it explains anything. The word posits two stories (a base below and something that rises from it) and then asks how the upper floor relates to the lower. To call phenomenal character emergent already grants that it stands *apart* from the physical goings-on and gets produced by them as a further fact, an extra the world lays on once the neurons fall into the right arrangement. That is the inner theater again, rebuilt in the vocabulary of complexity science. The identity claim denies the premise outright. Phenomenal character does not *arise from* representational content of the right embodied, world-directed kind; it *consists in* it. One fact under two descriptions — not a ground floor and a tenant upstairs. So the right response refuses the question rather than answering it. "Is consciousness emergent?" asks how one thing gives rise to a second thing, and we deny there are two things to relate.

Still, the word will not vanish by fiat, and it keeps a respectable use worth preserving. Distinguish the two senses it runs together.[30]

Weak emergence runs everywhere and costs nothing. Wetness, temperature, the lattice of a traffic jam, the wheeling of a starling flock: higher-level regularities wholly fixed by their physical base, deducible in principle from below once the identities and bridge facts are in hand — a posteriori

deducibility, not derivation from the bare physical description, a distinction Chapter 11 keeps sharp — merely too tangled to compute in practice. A physicalist pockets these without a flinch. If "consciousness emerges" means only this, fine — but then the word carries no anti-physicalist charge, and it has done none of the work its users wanted from it.

Strong emergence supplies the dangerous sense. Here the higher-level property counts as genuinely novel: not deducible even in principle from the complete physical story, and endowed with its own fundamental causal powers reaching back down upon the base. British emergentism in modern dress,[31] it amounts to dualism in a lab coat, which is why Chalmers parks his "naturalistic dualism" in the neighborhood, and why the panpsychist keeps arriving at the same address from the opposite direction.[32] Grant strong emergence for phenomenal character and you have granted that felt quality belongs among the fundamentals — the one thing the physicalism defended here exists to deny.

A deeper reason to distrust the word comes from Tim Crane: examine what an "emergent property" actually amounts to, and the line between full-blooded emergentism and respectable non-reductive physicalism refuses to stay drawn.[33] The vocabulary offers no safe middle berth between the identity claim and the second realm; it offers a fog in which the two keep trading coats. And a cautionary case sits close to home. Tye, in the very book that launched the modern representationalist program, recoiled from the thought that phenomenal consciousness had merely "emerged from the physical world" with no explanation of its emergence whatsoever — such brute emergence, he held, would be tantamount to magic, and he found it very hard to believe.[34] The instinct was right. Yet Tye, as Chapter 10 watched, went on to embrace panpsychism, seating felt character among the fundamentals after all. The moral runs subtler than "he grew careless with a word," and cuts deeper. Rejecting strong emergence will not save you if you keep the framing that forces the choice: *either* felt character arose from the wholly non-phenomenal *or* it sat among the fundamentals all along. Refuse the first horn while keeping the dilemma, and the second lies in wait. The identity claim refuses the dilemma itself: it denies there are two things, a base and an arising, to be related in either direction.

Which leaves the real diagnosis. What people hope *emergence* will bridge is no seam in the world where a new property bolts on. It is the gap in our concepts (the felt distance between a third-person account and the first-person quality it describes), and that gap, as Section 11.3 argued, lodges in us, not in nature. The identity claim does not explain how consciousness emerges. It explains why nothing has to. Ask the question the other way (how does matter *become* conscious?) and the answer holds its shape: it does not become anything.

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 epis-

temic 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, practical interest, and welfare. 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 identity claim makes the result more severe, not less. If phenomenal character consists in world-involving representational content of the right kind, then a duplicate with no content has no phenomenal character. The lights are, for that first instant, off. This is phenomenal externalism without a cushion: molecule-for-molecule duplication of the current body does not duplicate experience when the relations that fix content are absent. Block's protest identifies the cost accurately; it does not expose an internal contradiction. An intrinsic mental paint would spare Swampman, but Chapter 6 has already declined it.

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 and welfare. 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 hierarchical 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 welfare 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 welfare 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.

When you look at a ripe tomato in good light, your visual system tokens a state representing the tomato as *red*. The phenomenology of the case (and here I make a controversial assumption I will return to) presents redness as a vivid, unified, intrinsic-seeming property apparently sitting on the tomato's skin. According to the physics, the tomato has no such property. Surface reflectance profiles interact with incident light, the light with your photoreceptors, the signal with downstream neural circuitry. Somewhere in that cascade, your visual system produces an experience as of redness in this thicker sense. The property your experience attributes to the tomato does not, on a defensible physicalist view, exist out there waiting to be picked up. Your visual system reliably represents the tomato as having a property it does not have.

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.

The conclusion sits poorly with the phenomenology it was supposed to explain. We do not seem to experience the tomato as having a complicated relational surface property. We seem to experience it as red in a way that points beyond any relational story. To deny that our color experience represents anything over and above the discriminable surface property is to flatten the phenomenology into its functional shadow. Strong representationalism, which holds that phenomenal character consists in representational content, needs an account of content rich enough to *include* whatever in color experience goes beyond reliable tracking. Michael Tye saw this clearly enough to spend a whole book defending the view that color experience represents real, mind-independent properties of surfaces, because the strong representationalist's commitment to content-realism leaves no other option once the Galilean intuition is in play.[45] Whether one finds Tye's color realism convincing or not, his project illustrates the pressure: tracking semantics alone, without further substantive commitments about which properties experience represents, cannot underwrite the phenomenology.

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.

Such a story exists, and its outline is enough to put to work here. It involves the relational, world-directed structure of perceptual states; the role of recognitional capacities that go beyond what selection alone fixed; and the contribution of conceptual and inferential context to what a given perceptual state means *for the perceiver*. Dretske, in *Naturalizing the Mind*, distinguishes the merely indicating properties of a state from its full representational properties, and allows the two to come apart in ways that selection-history alone cannot adjudicate.[48]

Run the mechanism on the hard case, not just the easy one. What fixes content beyond ratification is chiefly the consumer side: the state feeds capacities that treat it as *red* in the thick sense — matching surfaces across changes of light, sorting by sameness of look, expecting the strawberry to match the tomato — and content follows what its consumers *have used* it as: the accumulated record of recruitment, correction, and stabilization in the system's own history, not the bare present disposition. (The distinction matters here, and B.9 collects it: at $t = 0$ Swampman has the consumers and none of the record, which is why present dispositions alone hand him no content.) That gives systematic mismatch somewhere to stand: were the world to lack the thick property, the state would still say it, the saying having been fixed downstream of the tracking. And the book should show its own hand here rather than leave the question open. Chapter 6 took its stand: colors are real — reflectance-types, out on the surfaces — so by this book's lights the tomato *satisfies what the experience actually attributes*: the reflectance-type, presented non-relationally, correct at the type's grain even where a perceiver's fine tuning misses (B.5 prices that concession). The systematic case then stays what Mendelovici needs it to be and no more: a live possibility, pressing on tracking theories, not a settled fact about human vision. Nor does the pressure need color for a home. If any register runs systematically rich, the affective one is the likelier candidate — a standing anxiety presents harmless mornings as menacing, attributing what the scene does not instantiate (B.6 above; Chapter 7 gives moods their content) — and Mendelovici's modal point never needed an actual case in any register: a theory that makes standing mismatch impossible *in principle* is refuted by the possibility, not the incidence. The richer story's yield is conditional but genuine. If the phenomenology anywhere outruns the physics, the framework no longer defines the mismatch away.

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?

It can. Fintan Mallory argues that backpropagation meets a consumer-based teleosemantic standard of selection: successive parameter configurations reproduce relevant structure under a lawful update rule, loss differentially preserves changes, and internal producer-consumer mechanisms thereby acquire proper functions ("Teleosemantics for Neural Word Embeddings," *Mind & Language* (2026), doi:10.1111/mila.70037). On his analysis, even a simple word2vec model contains intentional signs with original contents.[51] The claim cannot be dismissed by saying the model has no evolutionary past. Training is a real causal history of the artifact. Checkpoint selection belongs directly to the training lineage; selected functions may pass by causal descent to producer-consumer mechanisms in a deployed model; the deployed token has not thereby selected itself. This can count as the artifact's own selection history and can explain why its internal states have the causal roles they do.

The semantic proposal concerns a specific grain. In Section 5.1, Mallory argues that an embedding represents the fact that an actual word-type occurs in a particular context and directs its consumer to produce a distribution over context words. The output gets compared with distributions in the data; discrepancies drive changes to the producer and consumer mechanisms. Word-types and linguistic distributions are themselves features of the world. This account supplies more than a useful reading of numerical geometry: the proposed mapping helps explain the mechanism's training and subsequent use. Whether those relations fix accuracy conditions not exhausted by an interpreting practice remains the question of originality at this limited grain.

The permutation does not settle that question. Relabel every word index and coordinate the changes to the input and output matrices: the same actual word can still lead through the mechanism to the same distribution over actual context words. The coordinates change while the candidate content-fixing relation survives. Change the external dictionary without preserving that relation, and a candidate content-fixing relation changes too. Neither case shows that contextual content depends wholly on an interpreter's assignment. Mallory's own discussion makes the distinction: the mathematics alone does not force a probability-distribution reading, but the comparison with actual distributions and the history of correction give that reading an explanatory basis. His dictionary observation bears on why the embedding need not represent its proximal one-hot code; it does not withdraw the proposed content about words and contexts.

Biological function also depends on an external history, and human learners inherit semantic standards. Neither causal nor semantic inheritance disqualifies original content. Word2vec's vocabulary, corpus, task, and loss descend from human practices; that ancestry does not establish that those practices exhaust its states' accuracy conditions. The general criterion remains: history and producer-consumer use must fix correctness beyond an interpreting practice's assignment. The book leaves word2vec originality at the linguistic-distribution grain unsettled. Nor can a failure at another grain decide it. A system can represent facts about the word *maple* without

referring to the maple outside the window. Holding the corpus fixed while that tree changes tests the latter claim, not the former.

Whole-system attribution requires a further relation, but not a second act of semantic creation. A larger system understands by means of historically shaped states when reliable consumer use coordinates memory, correction, inference, conflict resolution, learning, planning, and action across mechanisms and timescales. A single monitoring loop bolted on for the occasion does not establish that economy; nor does causal embedding by itself. Flexible, counterfactually integrated use provides the positive evidence. An automated factory, distributed robot, or trained model might pass this test in a bounded domain, and the account should accept that result rather than protect a human monopoly. Reciprocal self-maintenance would add autonomy and a practical stake. It would not convert derived content into original content or manufacture understanding from otherwise unintegrated states.

The assessment carries its own revision conditions. Evidence that an artificial lineage's history and producer-consumer use fix accuracy conditions not exhausted by an interpreting practice would support original content at the relevant grain. Evidence that the proposed standard remains wholly practice-dependent would support a derived verdict. Embodied reinforcement learning, autonomous robotics, and artificial-life systems supply possible cases, not automatic victories or the only eligible routes. If a deployed system integrated content across flexible learning, inference, correction, planning, and action, the case for whole-system understanding would strengthen. If it also secured resources, repaired or replaced components, and preserved the organization realizing it, the case for autonomy and practical stake would strengthen separately. Manufacture disqualifies none of these achievements. Word2vec supplies a serious account of mechanism-level linguistic content while leaving its originality unsettled; that limited content claim establishes neither distal singular reference nor whole-system understanding, subjecthood, consciousness, autonomy, or welfare.[51a]

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 the gap Part Two spent itself closing opens again from the far side. 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.) Chapter 11 deflates the hard problem by relocating the gap into our concepts and closing it with the identity claim. Illusionism deflates harder. It says the gap guards nothing, because phenomenal consciousness — the felt glow the hard problem asks physics to explain — does not exist. Keith Frankish, the position's most forthright advocate, distinguishes weak illusionism, on which introspection misleads us about some features of experience, from strong illusionism, on which phenomenal properties themselves are an introspective illusion: experience seems to have a felt character, and the seeming exhausts the phenomenon.[53] Daniel Dennett had quined the philosophers' fourfold qualia decades earlier without denying the discriminations, pains, affects, memories, reports, and action-guiding patterns that ordinary experience talk tracks. Frankish turned the stronger denial of phenomenal properties into a program with its own research question — the illusion problem: explain, in functional terms, why experience seems phenomenal when nothing phenomenal exists.[54]

The kinship deserves stating before the quarrel. Strong illusionism and the identity claim agree on nearly everything a dualist cares about: one world, causally closed; no irreducible phenomenal residue; the hard problem answered by explaining the appearance of a problem rather than by bridging an ontological gulf. Chalmers's meta-problem names the shared workspace — why do creatures like us judge there to be a hard problem at all?[55] The illusionist answers: because introspection systematically misrepresents inner states as phenomenal. The identity claim answers: because acquaintance and description deliver one process under concepts that share no a priori route. Same explanandum, opposite verdicts on the datum.

That verdict makes the whole quarrel, and it matters where the illusion is supposed to sit. What illusionism debunks is an inner item with intrinsic qualitative properties — mental paint, the private glow. This book spent Part Two arguing that no such item exists to be found: look for the redness and you find the tomato. The illusionist reads that failure as exposure — introspection

promised a glow and could not produce one. The identity claim reads it as disclosure: introspection never presented an inner item, because experience presents the world. Where the illusionist calls the seeming of phenomenality an introspective error, the book holds that the seeming just is world-presenting content under a recognitional mode — real, physical, and exactly as it seems, once "seems" stops being read as a promissory note for paint. Nothing gets debunked, because nothing was promised except a presented world, and the world was delivered.

The cost accounting then favors the identity claim. Strong illusionism pays a Moorean price no theory should pay while an alternative remains standing: it must classify the one thing each of us cannot doubt — that seeing red differs from seeing green, that the ankle hurts — as a mistake about there being anything it is like at all. Frankish accepts the bill and argues that the intuitive resistance is itself predicted by the view. François Kammerer, the position's most careful developer, has named explaining that resistance the hardest aspect of the illusion problem and built a solution to it.[56] Grant the ingenuity. A theory that must work this hard to explain why everyone, including its defenders at dinner, finds it unbelievable might instead conclude that the datum survives — and the identity claim keeps the datum at no extra ontological charge: the same physicalism, the same closed causal order, with the felt difference between red and green preserved as a difference in world-presenting content. Kammerer presses back that any acquaintance-based defense of the datum invites the same debunking in turn. Chapter 11′s reply to its first objection holds here too: explaining a recognitional capacity in physical terms discredits its deliverances only if the explanation shows them insensitive to their subject matter, and on the identity claim the opposite holds — phenomenal judgments are caused by the very states they concern. The debunking arrow needs an orphaned judgment, and the orphaning belongs to the dualist's picture, not this one.

Illusionism looked into the empty theater and declared the show a hallucination. The book's verdict runs otherwise: the show is real, and it was never playing indoors.

B.17 Higher-Order Theories: The Monitor and the Missing Target

(Deep-dive from Chapter 6, The Identity Claim.) Chapter 6, note 7b states the book's two-edged case against higher-order theories in one paragraph; the exchange deserves full length, since the higher-order approach stands as first-order representationalism's most serious physicalist rival. David Rosenthal's version runs: a mental state becomes conscious when the subject has a suitable higher-order thought — a thought, arrived at non-inferentially, to the effect that one is in that state. Qualitative character belongs to the first-order state, located in a quality space of discriminable properties; consciousness of it arrives with the monitor. Hakwan Lau and Rosenthal have assembled an empirical case: prefrontal and parietal activity tracks visibility and confidence in ways first-order sensory strength does not.[57] The theory earns its standing by explaining something every view must explain: why some content-bearing states are conscious and others — blindsight discriminations, subliminal primes — are not.

Rosenthal has a reply to the inner-audience charge, and it deserves its hearing. The higher-order thought is not perception, not an inner eye, not a spectator. It is an ordinary intentional state whose presence constitutes, rather than observes, the first-order state's being conscious. Take the reply at full strength: no theater has been re-staged, because nothing watches anything; there is only a state, and a thought about it.

The misrepresentation problem survives that reply in full force. Higher-order states are representations, and representations can misfire. Let the monitor mischaracterize its target — represent a red-presenting state as green-presenting — or fire with no target at all. What is it like for the subject then? Say character follows the monitor, and the first-order state drops out: an empty higher-order thought yields full-blown phenomenal consciousness with nothing at the ground floor, and the world-presenting content that Part Two showed carrying every phenomenal difference has gone idle inside the theory of the very phenomenon it was supposed to constitute. Say character follows the monitored state, and the higher-order thought stops being constitutive: the first-order state carries its character, the monitor merely notices, and the view has become the first-order view under a new letterhead. Ned Block pressed the dilemma to the conclusion that the approach is defunct; Rosenthal's reply takes the first horn.[58] That reply need not predict that introspection reveals the higher-order thought: the thought may remain unconscious, and awareness of it would require another order. The comparative cost lies elsewhere. Targetless or erroneous higher-order content can fix what it is like while first-order world-presenting content becomes dispensable, reversing the explanatory order the transparency and hard-case arguments support.

The first-order alternative pays the higher-order theory's explanatory debt without the extra layer. What separates conscious from unconscious content, on the identity claim, is poising: the availability of world-presenting content to the systems that form beliefs, fix desires, and guide action. Blindsight content moves behavior and never reaches that economy; the primed word tilts a choice and never stands available for report. No second state need certify the first — the difference lies in the content's position within one cognitive organization, not in whether a monitor got hired to watch it. Uriah Kriegel's self-representationalism, which folds the monitor into the state itself, buys unity at the price of the same question one door down: the part of the state that represents the rest can still misfire, and the dilemma re-enters through the seam.[59] The identity claim keeps one layer, poised, and lets the world do the presenting. Nobody watches the watcher, because the job was never posted.

B.18 Integrated Information Theory: Integration Without a World

(Deep-dive from Chapter 20, note 2, 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 their interaction, used to predict sensory input, minimize prediction error, and guide perception, learning, and action through precision-weighted belief updating	Attention may be weights, not inference; a hierarchy may be a stack of layers; prediction error may be a training signal, not a perceptual state	Raise confidence by showing that the same latent variables, precision dynamics, and error profiles govern perception, learning, and action across tasks and timescales; lower it if the system tracks the world well while these signatures are absent
AST	A model of attention that represents what is being attended to, why, and with what effects, and uses that model to control processing	Attention may be a learned bias, not a control state; a report of attention may be a trained behavior; a system may track salience without representing attention	Raise confidence when attention-like control is causally necessary for flexible behavior, and when the system's reports about attention track its actual processing; lower it if attention can be removed without loss of control
Agency	Persistent goals, action selection, and learning from consequences, with behavior organized around maintaining viability or pursuing objectives	Goal-directedness may be a behavioral pattern, not an internal state; reinforcement may shape outputs without creating agency	Raise confidence when the system flexibly pursues goals across novel contexts, updates from consequences, and resists interventions that should block its objectives; lower it if apparent goals are brittle, externally imposed, or absent under pressure
Self-model	A representation of the system as an entity with boundaries, capacities, and persistence, used to regulate behavior	Self-reports may be trained outputs; a self-model may be a prompt artifact; a system may use "I" without representing itself	Raise confidence when self-representation is causally load-bearing, tracks the system's actual properties, and supports appropriate generalization; lower it if it is confabulated, inconsistent, or dissociable from behavior

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 count once 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:

> 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. It 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, 2006), 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 rejoinder the text presses — that global color eliminativism undercuts the perceptual contact it presupposes — see Alex Byrne, "Color and the Mind-Body Problem," *Dialectica* 60, no. 3 (2006): 223–244, esp. 227–28, and Alex Byrne and David R. Hilbert, "Color Realism and Color Science," *Behavioral and Brain Sciences* 26, no. 1 (2003): 3–21, esp. secs. 2–3. 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 concession-and-reply rehearsed in the text (the represented property is the coarse-grained reflectance-type, so small, stable misrepresentation of its fine grain by most perceivers is a prediction of the view, not an embarrassment to it), 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 conditions 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 text's objection — that an irreducible worldly color reinstates, for color alone, precisely the explanatory gap the book refuses elsewhere — restates the standard physicalist complaint that primitivism leaves color metaphysically and causally dangling.

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; the conscious state's visual mode supplies the phenomenality. 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. 1 (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. The weak/strong distinction has become standard since Chalmers drew it sharply: weak emergence names high-level facts unexpected from, yet deducible in principle from, the base; strong emergence names facts "not deducible even in principle" from it. See David J. Chalmers, "Strong and Weak Emergence," in *The Re-Emergence of Emergence*, ed. Philip Clayton and Paul Davies (Oxford: Oxford University Press, 2006). Searle drew the same line in 1992 in his own vocabulary: a feature is *emergent1* when its existence must be explained by the causal interactions among a system's elements — solidity, liquidity, and, on his view, consciousness — and *emergent2* when it additionally has causal powers that cannot be explained by those interactions. He endorses the first and rejects the second: "on my view consciousness is emergent1, but not emergent2." John R. Searle, *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992), ch. 5, at 111–112. Biological naturalism's approving talk of consciousness as a causally emergent property of systems is therefore weak emergence throughout — emergent1 maps onto the weak sense, emergent2 onto the strong — and concedes nothing to the doctrine this section rejects.

31. The lineage runs through Mill, Samuel Alexander, and C. D. Broad; the canonical history and post-mortem is 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, esp. sec. 6. The contemporary revival of strong emergence for consciousness is the same doctrine recast for the mind–body case.

32. For the property-dualist moral drawn from treating consciousness as, at best, strongly emergent, see Chalmers, *The Conscious Mind* (New York: Oxford University Press, 1996), esp. ch. 4 — the locus of "naturalistic dualism." The panpsychist reaches the same neighborhood from the far side: holding that phenomenal character could not strongly emerge from the wholly non-phenomenal, Goff concludes it must instead be fundamental — Philip Goff, "How Exactly Does Panpsychism Help Explain Consciousness?", *Journal of Consciousness Studies* 31, no. 3–4 (2024): 56–82. Either route seats felt character among the basic features of the world; that shared conclusion, not the direction of travel, is what the present view rejects. The book's reply to the panpsychist route runs in Chapter 10.

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 turn to panpsychism (*Vagueness and the Evolution of Consciousness*, 2021, ch. 4) lands him on the other horn of the same dilemma, making phenomenal character fundamental rather than emergent. The book's reply to that conversion — the epistemic-slide diagnosis together with the idle-wheel objection — appears in Chapter 10.

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, 2006), 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.

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, practical interest, and welfare. 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, practical interest, and welfare. (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), esp. chaps. 3–4. Tye defends the view that color experience represents real, mind-independent surface properties — a content-realism the strong intentionalist is pushed toward once the Galilean intuition (that color phenomenology presents an intrinsic, non-relational property) is in play. The book is the most sustained attempt to make tracking-style content rich enough to underwrite color phenomenology; whether it succeeds is exactly what the section leaves open.

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 opening the main text exploits to let the frog's state mean *fly* in a sense richer than what selection alone ratified.

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 leaves originality at this grain unsettled: human ancestry of the vocabulary, corpus, task, or loss alone does not establish that interpreting practice exhausts the relevant accuracy conditions. 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.

51a. 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), 560–585, 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. The weak/strong distinction is Frankish's own (sec. 1): weak illusionism concedes phenomenal consciousness while holding that introspection misrepresents some of its features; strong illusionism denies that experience has phenomenal properties at all, replacing them with *quasi-phenomenal* properties — physical properties that introspection misrepresents as phenomenal. Frankish argues that only the strong form dissolves the hard problem; the weak form inherits it. The special issue containing the paper, with commentaries, was reissued as *Illusionism as a Theory of Consciousness*, ed. Keith Frankish (Exeter: Imprint Academic, 2017).

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," secs. 1–2), 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 follows Chapter 11′s first objection-and-reply: physical explicability of a recognitional capacity discredits its deliverances only where the explanation orphans them from their subject matter.

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. Block's argument turns on the empty and erroneous higher-order thought: if consciousness is constituted by the higher-order state, a targetless monitor should yield conscious experience with no first-order state at all, which Block takes to reduce the view to absurdity. Rosenthal's reply accepts the consequence — what it is like for the subject is fixed at the higher-order level, target or none. The entry's transparency argument prices that acceptance: it makes first-order world-presenting content inessential to phenomenal character, against the introspective datum Chapter 4 established.

59. Uriah Kriegel, *Subjective Consciousness: A Self-Representational Theory* (Oxford: Oxford University Press, 2009). Kriegel's view makes a conscious state represent the world and, in the same stroke, itself, eliminating the separate monitor and with it the possibility of a *wholly* targetless higher-order state. The entry's point is that the misrepresentation structure survives internalization: the self-representing component can still mischaracterize the world-representing component, and the theory must again say which component phenomenal character follows.

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): 529–566, 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. The DOI and publication details were verified against the journal site. 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.

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 (1225–1274), the Italian Dominican friar who taught at Paris and Naples, built the most systematic philosophy of the European Middle Ages. His synthesis of Aristotle with Christian theology, worked out above all in the *Summa Theologiae*, shaped centuries of thought on nearly every question philosophy asks. Philosophy of mind remembers him for something more particular: an account of intentionality that anticipated distinctions the field would only rediscover in the nineteenth century.

Major Works

- *Summa Theologiae* (c. 1265–1274) — the comprehensive theological synthesis; the Treatise on Human Nature (Ia, qq. 75–89) addresses the soul, intellect, and perception
- *Quaestiones Disputatae de Anima* (c. 1268) — detailed treatment of the intellect and the soul's relation to the body
- *Commentary on Aristotle's De Anima* (c. 1268) — the systematic engagement with Aristotle's psychology that grounds Aquinas's own account

Key Positions

Esse naturale and esse intentionale. Aquinas's central contribution to the philosophy of mind is the distinction between *esse naturale* (natural existence) and *esse intentionale* (intentional existence). The form of a stone exists *naturally* in the stone; the same form can exist *intentionally* in the mind of someone who perceives or thinks of the stone. When the eye takes on the form of a colored thing without taking on the matter (the eye does not turn red when it sees red), the form is present intentionally. The distinction helps explain how a form can be present in cognition without the knower literally acquiring the object's matter. It belongs to the medieval background of representation, but the explicit distinction between a representation's vehicle and content emerges later, especially in Henry of Ghent and Duns Scotus.

Hylomorphism applied to mind. Following Aristotle, Aquinas held that the soul is the *form* of the body: not a separate substance lodged within it but the organizing principle that makes a living body what it is. Sensation, for Aquinas, is an act of the whole animated body, not of an immaterial mind peering through physical sense organs. This grounding of perceptual activity in the body is the pre-Cartesian alternative to the disembodied mind.

Intellectual operation and the subsisting intellect. Aquinas held that sensation requires a bodily organ while the intellect's grasp of universal forms does not, and on this basis he argued that the rational soul *subsists*: it can exist apart from the body. Even then, he insisted, a separated soul remains incomplete; on his view the human being is the body-soul composite, not the soul alone.

Abstraction and intelligible species. In knowing, the intellect abstracts the *intelligible species* from the *phantasm* (the sense image produced by perception). The intelligible species is that *by*

which the intellect knows, not that *which* it knows — an early formulation of the distinction between the instrument of cognition and its object. To know the tree is to know the tree *through* a form, not to inspect an inner picture of it.

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 an Oxford philosopher who died at forty-eight, leaving behind two short posthumous books and a permanent change to the philosophy of language and perception. He treated conceptual confusion the way a surgeon treats infection: precisely, without sentiment, and without mercy. His *Sense and Sensibilia* (1962), reconstructed from lecture notes, applies that method to the argument from illusion, the engine of indirect realism. The argument doesn't survive the operation.

Major Works

- *Sense and Sensibilia* (1962, posthumous) — dismantling the argument from illusion and the sense-datum theory
- *How to Do Things with Words* (1962, posthumous) — the performative/constative distinction and speech act theory
- "Other Minds" (1946) — knowledge, evidence, and the problem of other minds
- "A Plea for Excuses" (1956) — the method of ordinary language philosophy demonstrated on moral psychology
- "Ifs and Cans" (1956) — freedom and determinism through attention to ordinary usage

Key Positions

Dismantling the argument from illusion. Austin's *Sense and Sensibilia* takes the argument from illusion apart with characteristic care. The argument runs: sometimes things appear other than they are (a straight stick looks bent in water); in such cases we cannot be directly perceiving the physical object (what we see is bent, and the stick is not); so in all cases, even veridical perception, what we directly perceive must be something other than the physical object — a sense datum, an appearance, an inner item. Austin's response: the argument is riddled with equivocations and relies on a tendentious use of "appears," "real," and "directly." The phrase "directly perceive" does no work except to introduce the conclusion as a premise. The claim that illusion cases show we are never perceiving the physical world confuses cases of misperception with genuine perceptions.

Ordinary language as a philosophical tool. Austin's methodology: philosophical puzzles frequently arise from philosophers mishandling ordinary language — using terms in unfamiliar ways while relying on the ordinary ways to generate intuitions. The remedy is systematic attention to how words actually work in the contexts where they work well. This is not conservatism about language; it is an empirical method. The way ordinary speakers distinguish appearing from being, real from unreal, direct from indirect, carries information about the actual structure of perception that philosophical theorizing tends to destroy.

"Real" as a trouser word. Austin analyzes "real" not as a descriptive predicate but as what he calls a trouser word: it gets its sense by ruling out specific contrasts in context. A "real" duck rules out a toy, a stuffed exhibit, a decoy. A "real" pain rules out a fake, a hypochondriac's complaint. "Real" wears the trousers: the negative, what is being contrasted, does the semantic work. This analysis dissolves many puzzles about perception: asking whether we really perceive the external world requires specifying what the alleged not-real perception would be.

Speech act theory. Austin's other major contribution: the distinction between constative utterances (which describe states of affairs and can be true or false) and performative utterances (which *do* something — "I promise," "I hereby declare," "I name this ship"). The constative/performative distinction opened into the full taxonomy of speech acts (locutionary, illocutionary, perlocutionary) developed further by Searle.

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 Bender and Alexander Koller

Emily M. Bender (b. 1973) is an American computational linguist, Professor of Linguistics at the University of Washington and one of the most prominent skeptics of the claim that large language models understand language. Alexander Koller is a computational linguist at Saarland University in Germany. Together they wrote the 2020 paper whose central thought experiment, an eavesdropping octopus, put the sharpest recent argument about form and meaning into circulation. Neither is a philosopher of mind by trade; the field argues with their octopus anyway.

Major Works

- E. M. Bender & A. Koller, "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data" (*Proceedings of the ACL*, 2020) — the octopus thought experiment
- E. M. Bender, T. Gebru, A. McMillan-Major & S. Shmitchell, "On the Dangers of Stochastic Parrots" (2021) — the influential critique of large language models

Key Positions

The form/meaning distinction, made precise. Bender and Koller define *form* as the observable structure of language (the marks and sounds and their statistical arrangement) and *meaning* as the relation between those forms and something outside them: communicative intent, and the world the forms are about. Their thesis is stark: "a system trained only on form has a priori no way to learn meaning."

The octopus thought experiment. Two people on separate islands send telegraph messages; a hyper-intelligent octopus taps the cable and learns to predict and mimic the traffic perfectly, never having seen an island, a coconut, or a bear. It can sustain the conversation — until one islander, facing a bear, asks for help building a weapon. The octopus, fluent in the *form* of the exchange and innocent of its *meaning*, has nothing to offer. Pattern over the signal never became grasp of the world the signal was about.

Stochastic parrots. In the 2021 paper, Bender and colleagues argued that a language model, however fluent, is a system for "haphazardly stitching together sequences of linguistic forms … according to probabilistic information about how they combine, but without any reference

to meaning." The phrase named a now-central worry about reading comprehension into fluent output.

The octopus and the Chinese Room. The octopus is a contemporary, information-age cousin of Searle's Chinese Room. Where Searle's man manipulates symbols by rule, Bender and Koller's octopus manipulates them by learned statistics, closer to how a real language model works, and it reaches the same wall: command of form, however total, never crosses into meaning, because meaning was never *in* the form to be extracted.

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 (b. 1942) has taught philosophy at New York University since 1996, after twenty-five years at MIT. His work ranges across consciousness, perception, and the foundations of cognitive science, and his distinctions have a way of becoming the field's standard equipment: nearly every current debate about consciousness sorts itself with his phenomenal/access divide.

Major Works

- "Troubles with Functionalism" (1978) — the absent and inverted qualia objections to functionalism
- "On a Confusion about a Function of Consciousness" (*Behavioral and Brain Sciences*, 1995) — introduces the phenomenal/access consciousness distinction
- "Mental Paint" (2003, in Hahn & Ramberg, eds., *Reflections and Replies*) — the anti-representationalist critique of transparency
- "The Mind as the Software of the Brain" — survey article on functionalism and its limits

Key Positions

Phenomenal vs. access consciousness. Block's most widely adopted contribution is the distinction between **P-consciousness** (phenomenal consciousness: what it is like) and **A-consciousness** (access consciousness: availability for reasoning, report, and the control of behavior). A state is A-conscious when its content is freely available to processes of reasoning and verbal

report; it is P-conscious when there is something it is like to be in it. These can come apart: Block's imagined "super-blindsighter" would have A-conscious access to visual information with no P-consciousness of it, and rich experience may be P-conscious yet exceed what its subject can report.

Mental paint: the anti-representationalist case. Block's "Mental Paint" is the central contemporary challenge to strong representationalism. His thesis: purely externalist representationalist accounts of consciousness fail to capture the intrinsic qualitative character of experience — the "mental paint" that colors experience from the inside. Against transparency, Block argues that there *are* features you can notice by introspecting that no difference in the represented world will account for. His strategy is to find phenomenal differences without representational differences:

- **Blurry vision.** When a sharp edge blurs, something in how things seem changes, yet (Block claims) you do not see the edge *as* fuzzy the way a watercolor edge is fuzzy — the change is in the experience, not in the represented world.
- **Pressure phosphenes.** Press on a closed eye: drifting blobs of light appear. No external object is being represented, making it hard to relocate the phenomenal character into worldly content.
- **Inverted Earth / inversion cases.** Scenarios designed to hold representational content fixed while phenomenal character differs, or vice versa, to show the two can come apart.

If even one such case establishes a phenomenal difference without a representational difference, strong representationalism's identity claim fails.

Absent and inverted qualia against functionalism. Block's earlier work presses the absent-qualia and inverted-qualia objections against functionalism: a system could satisfy every functional specification of a mental state while having no experience at all, or systematically inverted experience. His most vivid illustration is the homunculi-headed system of "Troubles with Functionalism": let the population of China, linked by two-way radios, realize the functional organization of a mind, and ask whether that vast system feels anything at all.

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 an Austrian philosopher and psychologist who held chairs at Würzburg and Vienna, trained Husserl and Meinong among others, and set the terms for virtually everything that followed in the philosophy of mind. His 1874 *Psychology from an Empirical Standpoint* contains the idea that has proven indestructible: every mental phenomenon is characterized by what he called the intentional inexistence of an object, along with reference to a content or direction toward an object. Later philosophy calls this directedness *intentionality* and treats Brentano's formulation as the modern statement of the mark of the mental.

Major Works

- *Psychology from an Empirical Standpoint* (1874) — intentionality as the mark of the mental; the founding text of act psychology
- *On the Origin of Our Knowledge of Right and Wrong* (1889) — moral philosophy grounded in intentional attitudes
- *Descriptive Psychology* (1982, posthumous) — further development of the classification of mental phenomena

Key Positions

Intentionality as the mark of the mental. Brentano's central thesis: every mental phenomenon (perceiving, thinking, believing, desiring, fearing) exhibits intentionality, the property of being directed toward an object. Physical phenomena, by contrast, do not point beyond themselves. The redness of the apple is just there; the perception of the redness points toward the apple. This directedness is what makes mental phenomena mental. Later debates about whether all mental states are intentional (what about moods? pains?) or whether some non-mental states are intentional (thermostats? viruses?) take Brentano's claim as their starting point.

Intentional inexistence. Brentano introduced the phrase *intentional inexistence* to describe the way mental states involve their objects: the object "inexists" in the mental act — exists within it. This was not a claim that mental objects fail to exist, but rather that they have a distinctive mode of existence as *immanent* to the mental act. The tree I perceive is present to consciousness in a way that ordinary material presence is not. The formulation invited a reading on which the object of perception is always an inner mental item — a reading later philosophy worked hard to undo, and one the older Brentano himself resisted.

The classification of mental phenomena. Brentano classified mental phenomena into three fundamental classes: presentations (*Vorstellungen*, the basic acts of taking something as an object), judgments (acts of affirming or denying something about an object), and phenomena of love and hate (the basic evaluative attitudes). Everything mental reduces to one of these. The classification

influenced Husserl's phenomenology and, indirectly, the taxonomy of propositional attitudes in contemporary analytic philosophy.

Act psychology. Mental states, on Brentano's view, are acts (things a mind *does*) rather than contents (things a mind *has*). This emphasis on the active, dynamic character of mental life runs against the empiricist tradition of minds as passive containers of ideas, and it connects to the broader claim that mental states are inherently directed rather than self-contained.

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 (b. 1946) has been at UCLA for most of his career, working with a precision and patience that often makes the difficulty of the problems invisible until you try to solve them yourself. His 1979 paper "Individualism and the Mental" set up a thought experiment that the field is still unpacking: imagine a person who holds a slightly mistaken belief about the meaning of "arthritis." Hold that person's physical and internal states fixed, but change the linguistic community around them so that "arthritis" means something else there. Their belief changes content — which means mental content is not fixed by what is inside the individual. Anti-individualism. The implications are still working themselves out.

Major Works

- "Individualism and the Mental" (1979) — the arthritis thought experiment; the founding paper of social externalism
- "Other Bodies" (1982) — environmental externalism extended to thought about the external world

- *Origins of Objectivity* (2010) — a monumental account of how perceptual representation constitutes objective knowledge of the world
- "Perceptual Entitlement" (2003) — the epistemic ground of perceptual justification
- "Individualism and Self-Knowledge" (1988) — how anti-individualism and privileged self-knowledge can be reconciled

Key Positions

Anti-individualism (social externalism). Burge's arthritis case: a subject holds the belief "I have arthritis in my thigh." In fact, the subject only has arthritis in their joints; the belief is false. Now run the counterfactual: the same subject, unchanged inside, in a community where "arthritis" covers thigh inflammations as well. There the belief's content differs; it is a different belief. The content of mental states depends constitutively on the linguistic and social community in which the subject is embedded. Hold the individual fixed; change the community; change the mind. This extends Putnam's externalism about natural-kind terms into the territory of ordinary concepts, showing that the phenomenon is not restricted to scientific vocabulary.

Perceptual anti-individualism. Burge extends anti-individualism from conceptual content to perceptual content. The veridicality conditions of perceptual states depend not just on internal functional organization but on the causal-environmental history of the perceptual system — on how systems of its kind have typically interacted with the world, not merely on the proximal stimulus or the internal state.

Origins of Objectivity. Burge's most ambitious work: a systematic account of how perceptual representation constitutes objectivity — the ability to represent the world as it is independently of any particular perspective. Perceptual states have representational contents with veridicality conditions; these contents are constitutively anti-individualistic, shaped by the causal-environmental history of the system. Perception, Burge argues there, is the ground of our capacity for objective representation, preceding and grounding conceptual thought.

Individuation of mental states. Mental states cannot be individuated by their intrinsic features alone. Two individuals who are physically and functionally identical but embedded in different environments can have mental states with different contents. Wide content (content individuated by the environment) is ineliminable. This conclusion has significance for cognitive science, which often assumes narrow, intrinsically individuated content.

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 Chalmers (b. 1966), an Australian philosopher at New York University, co-directs NYU's Center for Mind, Brain, and Consciousness. He arrived in philosophy of mind via mathematics and cognitive science, and the combination shows: Chalmers brings a systematic, almost engineering-grade rigor to the most unsystematic of problems. He gave the field the phrase "the hard problem of consciousness," first in a 1995 paper and then at book length in *The Conscious Mind* (1996), and the debate has run on his framing ever since. Whether or not you find his conclusions persuasive, the problems he posed have not gone away.

Major Works

- *The Conscious Mind: In Search of a Fundamental Theory* (1996) — the hard problem, zombie argument, and property dualism developed in full
- "Facing Up to the Problem of Consciousness" (*Journal of Consciousness Studies*, 1995) — the original hard problem paper
- "Does Conceivability Entail Possibility?" (2002) — the modal epistemology behind the zombie argument
- "The Representational Character of Experience" (2004) — engagement with intentionalism and two-dimensional semantics
- "The Content and Epistemology of Phenomenal Belief" (2003) — phenomenal concepts and the knowledge argument
- "The Meta-Problem of Consciousness" (*Journal of Consciousness Studies*, 2018) — why we judge consciousness puzzling, posed as a tractable research program
- *Reality+: Virtual Worlds and the Problems of Philosophy* (2022) — virtual reality as a lens on metaphysics and consciousness

Key Positions

The hard problem of consciousness. Chalmers distinguishes the *easy problems* of consciousness (explaining attention, integration, reportability, the access of information to cognition) from the *hard problem*: explaining why there is something it is like to undergo these processes at

all. The easy problems are not trivially easy, but they are tractable in principle by the standard methods of cognitive science. The hard problem, Chalmers argues, is different in kind. Any functional or physical story leaves open why that story is accompanied by phenomenal experience. Even a complete neuroscience would leave the question "but why does it *feel* like something?" unanswered.

The zombie argument. Chalmers argues that philosophical zombies (beings physically and functionally identical to us, but with no phenomenal consciousness at all) are conceivable, and that their conceivability entails their possibility. If zombies are possible, then phenomenal consciousness is not fixed by physical facts alone, and physicalism fails. The argument turns on the inference from ideal conceivability to genuine metaphysical possibility, which Chalmers defends at length in "Does Conceivability Entail Possibility?" (2002) against the standard objection that conceivability guides possibility poorly.

Two-dimensional semantics. Chalmers develops a sophisticated framework to handle the epistemology of modal claims. Every expression has a primary intension (what it picks out in the context of utterance, roughly its epistemic content) and a secondary intension (what it picks out in all possible worlds, its metaphysical content). The zombie argument operates at the level of primary intensions: zombies are conceivable because no a priori analysis of physical concepts yields phenomenal concepts. The two-dimensional framework is the modal engine behind his anti-materialism.

Property dualism and panpsychism. Chalmers concludes from the zombie argument that phenomenal properties are not identical to, nor logically supervenient on, physical properties. He calls this "naturalistic dualism": two kinds of property, one physical and one phenomenal, in a single natural world. His later work takes panpsychism (the view that phenomenal properties are fundamental features of reality at every level) seriously as a candidate for explaining how the physical and phenomenal relate without positing a brute unexplained correlation.

The meta-problem. Chalmers' 2018 paper "The Meta-Problem of Consciousness" turns his method on the problem itself: explain why we make the judgments we do about consciousness — why we report a hard problem in the first place. Unlike the hard problem, the meta-problem is functionally tractable, since reports and judgments are behavior. Chalmers offers it as a research program that should constrain every theory of consciousness, his own included; illusionists read it as the opening move in explaining the hard problem away, a reading he considers and declines.

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 automat-

ically 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, one of the founders of theoretical computer science and the doctoral supervisor, at Princeton, of Alan Turing. He invented the lambda calculus in the early 1930s, proved the undecidability of first-order logic in 1936, and formulated the thesis, later joined to Turing's independent analysis, that fixes what we mean by "computable" to this day. His name sits on the single most cited, and most often misused, principle in the philosophy of computation.

Major Works

- "An Unsolvable Problem of Elementary Number Theory" (*American Journal of Mathematics*, 1936) — the lambda calculus and the first formulation of what became the Church–Turing thesis
- "A Note on the Entscheidungsproblem" (*Journal of Symbolic Logic*, 1936) — Church's theorem: there is no general decision procedure for validity in first-order logic
- "A Formulation of the Simple Theory of Types" (1940) — foundational work in higher-order logic

Key Positions

The lambda calculus. Church built a spare formal system in which everything is a function and computation is the rewriting of expressions by fixed rules. It turned out to capture exactly the same class of computable functions as Turing's machine (two utterly different formalisms converging on one boundary), and it remains the theoretical backbone of functional programming.

Church's theorem. Church proved that no algorithm can decide, for an arbitrary statement of first-order logic, whether it is valid. Together with Turing's contemporaneous result, this settled Hilbert's *Entscheidungsproblem* in the negative: some questions admit no mechanical decision procedure at all.

The Church–Turing thesis. The proposal that the *effectively calculable* functions (those computable by any idealized mechanical procedure) are exactly the ones definable in the lambda calculus, equivalently the Turing-computable ones. It is a claim about the extent of mechanical computation. It says nothing, on its own, about minds.

The thesis and its strengthenings. In its proper form the thesis bounds what can be mechanically computed. Later writers strengthened it: a "physical" or "strong" Church–Turing thesis holds that every physical process can be simulated by a Turing machine, and the Church–Turing–Deutsch principle extends the claim to physics as a whole. In philosophy of mind, the strengthened readings underwrite the idea that the mind itself is a computation. Neither strengthening is Church's own, and neither follows from the mathematics.

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 (b. 1962), a British philosopher, has held chairs at University College London and Cambridge and now works at Central European University. Lucidity comes in two kinds in this field: the kind that oversimplifies the problem, and the kind earned by finding out what you actually think. Crane writes the second kind. His *Elements of Mind* (2001) remains the clearest single-volume introduction to the problems this book engages, and his work on intentionality and consciousness has shaped the debate more than the citation counts show.

Major Works

- *Elements of Mind: An Introduction to the Philosophy of Mind* (2001) — a systematic overview of intentionality, consciousness, mental causation, and the mind-body problem
- *The Mechanical Mind* (1995; 3rd ed. 2015) — computers, functionalism, and the nature of mind for a general audience
- *The Objects of Thought* (2013) — a sustained account of how mental states can be directed toward nonexistent objects
- "Intentionality as the Mark of the Mental" (1998) — defense of the Brentano thesis against behaviorist and functionalist alternatives

- "The Intentional Structure of Consciousness" (2003) — argument that consciousness cannot be dissociated from intentionality

Key Positions

The intentional structure of consciousness. Crane holds that phenomenal character is determined by intentionality, but he rejects the content-only version often associated with strong transparency. A conscious perception has a subject–mode–content structure. What it is like depends both on what the state presents and on how it presents it — visually rather than auditorily, painfully rather than neutrally, fearfully rather than as a bare judgment. No non-intentional mental paint follows from that distinction.

Against the dissociation of consciousness from intentionality. Crane resists treating phenomenal consciousness as an intrinsic property detachable from intentional structure. Strip away content and mode and no free-standing phenomenal residue remains. His disagreement with content-only representationalists concerns the completeness of their analysis, not a defense of non-representational qualia.

The objects of thought. A challenge for any intentionalist: mental states can be directed toward things that don't exist — Sherlock Holmes, the fountain of youth, round squares. What is their object? Crane resists both the Meinongian answer (there are nonexistent objects, somehow) and the eliminativist answer (such thoughts are just confused). His solution: intentional objects are not a special category of entity at all. To specify a thought's intentional object is to say what the thought is about, not to inventory a thing; a Holmes-thought has an object in this schematic sense while relating to no real detective.

History of consciousness. Crane has written extensively on the history of philosophical theories of consciousness, tracing the trajectory from Brentano through Husserl, the analytic tradition, and contemporary philosophy of mind. That sense of where the problems came from is something this book shares.

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 (b. 1956) launched the modern scientific search for the neural correlates of consciousness. Crick, the British molecular biologist who had co-discovered the structure of DNA, spent the last quarter of his life on the brain; Koch, a German-American neuroscientist then at Caltech and later chief scientist at the Allen Institute, was his close collaborator and has carried the program forward. Together they made consciousness a respectable laboratory problem. Apart, their later paths (Crick's austere reductionism, Koch's turn toward Integrated Information Theory) map the two ways a science of consciousness can go.

Major Works

- Francis Crick, *The Astonishing Hypothesis: The Scientific Search for the Soul* (1994)
- Crick & Koch, "Towards a neurobiological theory of consciousness" (*Seminars in the Neurosciences*, 1990) — the founding manifesto of the NCC program
- Christof Koch, *The Quest for Consciousness* (2004), *Consciousness: Confessions of a Romantic Reductionist* (2012), *The Feeling of Life Itself* (2019)

Key Positions

The neural correlates of consciousness (NCC). Crick and Koch proposed setting the philosophy aside and asking a tractable question: what is the minimal set of neural events sufficient for a specific conscious experience? The strategy made consciousness an experimental target rather than a metaphysical impasse, and it has organized the field ever since.

The astonishing hypothesis. Crick's slogan — that "you, your joys and your sorrows, your memories and your ambitions … are in fact no more than the behavior of a vast assembly of nerve cells" — is uncompromising physicalism stated for a popular audience, with a deliberately provocative reductive edge.

Koch's later turn. Where Crick stayed an austere reductionist to the end, Koch came to hold that the NCC program, by itself, could not say *why* any neural process is accompanied by experience. He adopted Tononi's IIT and with it a graded, near-panpsychist view on which consciousness is far more widespread than the brain. In 2023 Koch publicly conceded a 25-year wager with David Chalmers that the neural mechanism of consciousness would be pinned down by then.

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 (b. 1965), a French cognitive neuroscientist, holds the Chair of Experimental Cognitive Psychology at the Collège de France and directs the NeuroSpin imaging center near Paris. With Jean-Pierre Changeux and Lionel Naccache he developed the Global Neuronal Workspace theory, the leading neuroscientific account of what happens in the brain when a piece of information becomes conscious. His work is a model of what the empirical science of consciousness can deliver — and a useful test of where that delivery stops.

Major Works

- *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts* (2014) — the global workspace theory for a general audience
- *Reading in the Brain* (2009) and *The Number Sense* (1997) — landmark work on literacy and numerical cognition
- Dehaene, Changeux & Naccache, foundational papers on the global neuronal workspace model of conscious access

Key Positions

The Global Neuronal Workspace. Dehaene's central idea is that the brain processes most information locally and unconsciously, and that a stimulus becomes *conscious* when it is amplified and broadcast across a network of long-range cortical connections, making it globally available to memory, language, attention, and report. Consciousness, on this account, is global information sharing — "the massive sharing of pertinent information throughout the brain."

Signatures of conscious access. Dehaene's lab identified reproducible neural markers that distinguish stimuli a subject reports seeing from ones processed unconsciously — chiefly a late, widespread surge of activity, the "global ignition," arriving around 300 milliseconds. This turned conscious access into something experimentally tractable.

A functional, broadcasting picture of mind. Global Neuronal Workspace theory proposes that information becomes consciously accessible when it is amplified and made globally available

to memory, language, reasoning, and report. This makes GNW a leading candidate account of the state-consciousness threshold. It does not by itself identify phenomenal character with broadcasting or settle why globally available content has the character it does.

Workspace versus IIT. Global Neuronal Workspace theory is one of the leading empirical theories of consciousness and a natural foil to Tononi's IIT; the two were tested against each other in a large adversarial collaboration. Where IIT begins from axioms about phenomenology and reaches toward a highly expansive distribution of consciousness, GNW begins with experimentally tractable access, ignition, and broadcast.

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) trained under Quine at Harvard and Ryle at Oxford, then spent five decades at Tufts University, where he co-directed the Center for Cognitive Studies. He argued from first book to last that consciousness, intentionality, and the self all yield to explanation by the natural sciences — and that the sense of a hard, irreducible mystery surrounding them is itself a philosophical artifact rather than a genuine discovery.

Major Works

- *Content and Consciousness* (1969) — his doctoral work; early case for a functionalist approach to mental content
- *Brainstorms* (1978) — collected essays on intentionality, the Intentional Stance, and folk psychology
- *Elbow Room* (1984) — compatibilism and the evolution of free will
- *The Intentional Stance* (1987) — the strategy of treating systems as if they had beliefs and desires, without committing to the reality of those states as inner objects

- *Consciousness Explained* (1991) — his major statement on consciousness: the Multiple Drafts model, heterophenomenology, the Cartesian Theater critique, and the self as a center of narrative gravity
- *Darwin's Dangerous Idea* (1995) — evolution as a universal acid that dissolves the case for any kind of "skyhook," including non-physical mind
- *Breaking the Spell* (2006) — religion as a natural phenomenon
- *From Bacteria to Bach and Back* (2017) — competence without comprehension; how evolution produces "know-how" without any inner knower

Key Positions

The Cartesian Theater and Multiple Drafts. Dennett's central target is what he calls the Cartesian Theater: the tacit assumption, found in both folk psychology and much cognitive science, that there must be a single place in the brain where "it all comes together," a screen on which experience is presented to an inner witness. Against this, he proposes the Multiple Drafts model: consciousness consists in multiple, parallel processes without any privileged moment of "becoming conscious" and without any central stage. There is no finish line, no inner display.

Heterophenomenology. Dennett's methodological proposal for studying consciousness scientifically without either dismissing first-person reports or treating them as authoritative. The heterophenomenologist takes subjects' reports seriously as data about what subjects *believe* their experience is like, while remaining agnostic about whether those beliefs accurately reflect any inner reality. The first-person perspective is treated from the outside, in the third person.

The Intentional Stance. Dennett treats belief-and-desire attribution as a predictive strategy: we adopt the intentional stance when interpreting a system as a rational agent makes its behavior intelligible. The stance is not therefore arbitrary or merely fictional. In his real-patterns account, robust and economical predictive success can disclose objective patterns in the system, even when no belief-shaped object sits inside it. What that success warrants, and at which level of organization, remains a further question.

Quining qualia. Dennett argues that nothing satisfies the traditional job description of qualia as intrinsic, ineffable, private properties directly apprehended by their subject. The convictions supporting that picture become phenomena for a science of consciousness to explain rather than evidence for mental paint. This makes him a major precursor of contemporary illusionism, though the later strong-illusionist claim that phenomenal properties do not exist should not simply be read back into every part of Dennett's account.

The self as a center of narrative gravity. Drawing on the metaphor of a center of gravity (a useful theoretical fiction that tracks real patterns without naming a particular object), Dennett argues that the self is an abstraction spun out of the stream of self-narrative the brain continuously produces. There is no inner owner; there is only the story, and the "self" is the fictional protagonist of that story.

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), educated by the Jesuits at La Flèche, spent most of his working life in the Dutch Republic and died in Stockholm in the service of Queen Christina of Sweden. The standard story names him the founder of modern philosophy, and for once the standard story holds. He invented analytic geometry (the Cartesian coordinates bear his name) and produced the most consequential reorientation of European thought since the late medievals: a turn inward, to the self-knowing mind, as the secure starting point from which the rest of knowledge would be rebuilt. Modern philosophy of mind as a distinct subject begins with him.

Major Works

- *Discourse on the Method* (1637) — the public manifesto of the new philosophy, published anonymously with three scientific appendices including the *Geometry*
- *Meditations on First Philosophy* (1641) — the systematic statement, published with six sets of Objections (Caterus, Mersenne, Hobbes, Arnauld, Gassendi, "Various Theologians") and Descartes' Replies
- *Principles of Philosophy* (1644) — the comprehensive textbook version, intended for the schools
- *Passions of the Soul* (1649) — the late account of mind-body union and the affective life; written for Princess Elisabeth of Bohemia
- *Treatise on Man* / *The World* (composed by 1633, withheld after the condemnation of Galileo; published posthumously in 1664) — the mechanistic physiology that grounds the dualism
- The *Correspondence with Princess Elisabeth of Bohemia* (1643–1649) — the most candid engagement with the interaction problem

Key Positions

Method of doubt and the cogito. In the *Meditations*, Descartes resolves to reject as if false anything that admits of the slightest doubt, in order to build knowledge from a foundation that cannot be shaken. Sense experience, mathematics, and even his own body all fall to the doubt; only the indubitable fact that *while he doubts, he thinks, and while he thinks, he exists* survives. *Cogito*

ergo sum: the thinker, in the act of thinking, knows itself to be. That becomes the foundation for everything that follows.

Substance dualism (the real distinction). In Meditation VI, Descartes argues that mind and body are two really distinct substances. The mind is a thinking, unextended thing (*res cogitans*); the body is an extended, unthinking thing (*res extensa*). Each can be conceived clearly and distinctly without the other, and what can be so conceived God can make exist apart. The argument grounds a metaphysical dualism in an epistemic capacity to conceive.

The mind better known than the body. Meditation II: the wax argument. A piece of wax, melted, transforms in every sensible property, yet I know it as *the same wax*. The continuity comes from the intellect, not the senses — and the lesson Descartes draws is that *the mind is better known than the body*. This is the epistemic priority claim that anchors the inner-theater tradition: I know my own mental contents more immediately and more certainly than I know any external thing.

Clear and distinct ideas as the criterion of truth. Meditation III. Whatever I perceive clearly and distinctly must be true — secured, in Descartes' argument, by the existence of a non-deceiving God whose existence is itself established by an a priori argument from the idea of God. The criterion underwrites the rest of the program: mathematics, the physical sciences, the existence of the external world.

Mind-body interaction and the pineal gland. In *Passions of the Soul* and the Elisabeth correspondence, Descartes locates the site of mind-body interaction at the pineal gland. How an immaterial mind moves a material body remains the canonical unresolved problem of his system — pressed sharply by Elisabeth and never satisfactorily answered.

Mechanism for bodies; animals as automata. Bodies, human and animal alike, are pure mechanism — complex assemblies of matter operating by mathematical laws. Animals, lacking minds, are nothing but automata. The mechanistic physiology of the *Treatise on Man* provides the physical half of the dualism.

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) spent nearly three decades at the University of Wisconsin–Madison before moving to Stanford and, late in his career, Duke. He worked with a philosopher's stubbornness on problems most of his contemporaries were happy to leave to psychologists and cognitive scientists, starting in epistemology and then building, from *Knowledge and the Flow of Information* (1981) onward, a naturalistic theory of representation. By the time *Naturalizing the Mind* appeared in 1995, that theory had converged with Michael Tye's on the claim at the center of this book: phenomenal character just is a certain kind of representational content, cashed out in Dretske's case in causal-informational terms.

Major Works

- *Seeing and Knowing* (1969) — perception and epistemic justification
- *Knowledge and the Flow of Information* (1981) — information-theoretic foundations for content and knowledge
- *Explaining Behavior: Reasons in a World of Causes* (1988) — mental causation and intentional explanation
- *Naturalizing the Mind* (1995) — the representational thesis; phenomenal character as representational content

Key Positions

Informational semantics. Dretske's foundational move: a mental state carries information about X when, under the relevant conditions, the state would not occur unless X were the case. This is the basic informational channel. On this picture, representation is a natural relation, causal and nomological, rather than something mind imports from outside nature. The resulting semantics is externalist: what a state is about depends on what it covaries with in the world, not on anything internal to the subject.

The representational thesis. In *Naturalizing the Mind*, Dretske argues that the phenomenal character of an experience (what it is like to see red, to feel a sharp pain) is the representational content of that experience: the properties it represents the world or body as having. Phenomenal properties are not inner objects or mysterious qualia floating free of representation; they are world-directed or body-directed representational properties of the experience itself. Dretske arrives at this view from the informational side; Michael Tye reached the same identity indepen-

dently that same year, in the theory he called PANIC (Poised, Abstract, Nonconceptual, Intentional Content). They converge.

The transparency of experience. When you introspect your experience of a ripe tomato, attention lands on the tomato's redness and roundness rather than on a second inner colored item. Dretske connects this transparency to representational naturalism: experience presents worldly or bodily properties through states whose representational functions depend on information-bearing relations recruited through design or learning history. A bare information channel is not yet a representation.

Mental causation and the causal relevance of content. Dretske works carefully on how intentional properties, the contents of beliefs and desires, can be causally relevant in a physical world. His answer: mental representations acquire causal powers by being recruited into functional roles; their content is causally relevant because the system was structured (by learning or evolution) to respond to that content.

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 (b. 1959), a British theoretical neuroscientist and Professor of Imaging Neuroscience at University College London, ranks among the most cited scientists alive. He built much of the statistical apparatus modern brain imaging runs on, and then proposed a single principle, the *free energy principle*, meant to explain why any self-organizing system, brain or cell or organism, persists at all. It is the most ambitious formalization of the predictive brain, and it carries both a deep affinity with the book's view of life and a deep temptation the book resists.

Major Works

- Statistical Parametric Mapping (SPM), voxel-based morphometry, dynamic causal modelling — the standard tools of human neuroimaging
- "The free-energy principle: a unified brain theory?" (*Nature Reviews Neuroscience*, 2010)
- A large body of work developing active inference and the free energy principle "for a particular physics"

Key Positions

The free energy principle. Friston's claim is that any system that maintains itself against dissolution must act so as to minimize a quantity (variational free energy) that bounds the surprise of its sensory states. To keep existing, a system must keep its expected and actual states in register. Perception, action, and learning all fall out as ways of minimizing this single quantity.

Active inference. On this picture a creature does not merely update its model to fit the world; it acts on the world to make the world fit its model. Perception and action are two routes to the same end, reducing prediction error, and an agent is a thing that does both in the service of staying alive.

The Markov blanket. Friston uses the statistical notion of a Markov blanket to formalize conditional-independence boundaries between internal and external states, mediated by sensory and active states. The framework can model organism-like boundaries, but blankets may occur at several nested scales; the mathematics alone does not uniquely decide which bounded organization counts as an organism, agent, or subject.

Scope and its critics. The free energy principle is predictive processing pushed to its most general and most metaphysical: read narrowly, a theory of biological self-maintenance; read expansively, a near-universal account of mind, life, and even physics. Critics divide sharply over which Friston they are reading, and over whether a principle that explains everything can pin down anything.

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) spent nearly his entire career at Princeton, and his range ran unusually broad: epistemology (inference to the best explanation, coherentism), ethics (the relativity of morality), philosophy of language, and philosophy of mind. In 1990 he published "The Intrinsic Quality of Experience," a short paper that became the hinge of the contemporary transparency literature and the starting point for almost every serious discussion of strong representationalism.

Major Works

- *Thought* (1973) — a functionalist account of mental content and meaning
- "The Intrinsic Quality of Experience" (*Philosophical Perspectives* 4, 1990) — the foundational transparency argument; the Eloise example; the three-objections-collapse
- *Change in View* (1986) — epistemology of rational belief revision
- *Reasoning, Meaning, and Mind* (1999) — collected essays spanning epistemology and philosophy of mind

Key Positions

The transparency argument and Eloise. Harman's central claim is that when Eloise attends to her visual experience of a tree, the colors and shapes she finds appear as properties of the tree and its surroundings, not as intrinsic paint on an inner experiential object. Introspection discloses the represented scene rather than a second colored bearer behind it.

The properties-of-object vs. properties-of-experience distinction. Behind the observation sits a diagnosis: qualia realism rests on a *confusion* — mistaking properties of the (possibly nonexistent) intentional object of experience for properties of the experience itself. The redness you are tempted to call the "quale" is the redness the experience *attributes to the tomato*. There is no second, inner red thing. Naming this confusion is what robs the qualia thesis of its air of obviousness.

Three objections collapsed at once. Harman's paper argues that three classic anti-physicalist objections (the appeal to direct awareness of intrinsic experiential qualities, Mary's Room, and the inverted spectrum) all collapse once the object/experience distinction is in place. The same transparency move that dissolves qualia also defuses the Knowledge Argument and blunts inversion scenarios.

Diaphanousness inherited from Moore. Harman's transparency thesis is the modern development of G. E. Moore's 1903 observation that the consciousness of blue is "diaphanous" — something we see through, not at. Harman turned a phenomenological remark into a philosophical argument with anti-qualia teeth. The genealogy runs Moore → Harman → Tye.

Psychophysical functionalism as the frame. Harman's transparency point was embedded in a defense of psychophysical functionalism — the view that mental states are individuated by their functional roles. The transparency datum does not by itself settle questions about what

fixes representational content, and Harman did not develop the externalist and embodiment dimensions that Tye would later add.

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 (b. 1947) is a British-Canadian computer scientist and cognitive psychologist whose work helped turn neural networks from a marginal research program into the dominant approach to artificial intelligence. His demonstrations that backpropagation could learn useful internal representations, followed by the AlexNet image-recognition breakthrough from his Toronto group, helped set off the modern deep-learning revolution. In 2024 he shared the Nobel Prize in Physics with John Hopfield for foundational work on machine learning with artificial neural networks.

Hinton began with a question about human minds rather than a plan to build commercial chatbots. He wanted to understand how brains learn representations without receiving a hand-written symbolic program. That research program has grown into a recognizable, if informally presented, philosophy of mind: connectionist about thought, naturalist and anti-Cartesian about consciousness, functional about mental-state attribution, and increasingly confident that contemporary AI systems possess genuine understanding and at least some forms of experience.

Major Works

- "Learning Distributed Representations of Concepts" (1986) — the family-tree model that learned semantic features and relations rather than receiving explicit symbolic rules
- "Learning Representations by Back-Propagating Errors" (with David Rumelhart and Ronald Williams, 1986) — the landmark demonstration of backpropagation's power to learn internal representations

- AlexNet (with Alex Krizhevsky and Ilya Sutskever, 2012) — the deep-network result that transformed computer vision and opened the floodgates for modern deep learning
- "The Forward-Forward Algorithm" (2022) — an alternative learning procedure whose closing section proposes "mortal computation"
- Public lectures and interviews (2023–2026) on learned meaning, machine understanding, subjective experience, consciousness, digital intelligence, and catastrophic AI risk

Key Positions

Connectionism: vectors before symbols. Hinton rejects the classical picture in which cognition fundamentally consists of explicit symbols transformed by formal rules. Brains evolved for perception and motor control before logic or language appeared, so he treats conscious symbol manipulation as a late achievement built on non-symbolic machinery. Inside a neural system, learned vectors of activity encode features and their interactions; symbols occur mainly at linguistic input and output. In his blunt 2017 formulation, those vectors of activity *are* thoughts. Formal reasoning matters, but intuitive and analogical cognition carries more of the load than classical AI allowed.

Meaning as learned relational structure. Hinton combines two familiar theories of word meaning: words take their significance from relations to other words, and words activate clusters of semantic features. A network learns both at once by converting words into high-dimensional feature vectors whose interactions support prediction, disambiguation, analogy, and inference. Text itself does not contain meaning on his account; a system understands a sentence when it constructs feature representations that fit together and can guide further cognitive work. Prediction does not compete with understanding, because learning to predict across varied contexts forces the network to extract the structure that makes those predictions possible. He treats one-shot inference of an invented word, reinterpretation of an ambiguous sentence, and explanation of layered jokes as evidence that current models have crossed from formal mimicry into understanding. Human beings and language models therefore understand in the same basic way, though humans learn more efficiently through multimodal engagement with the world.

Hinton's 2017 account also gave mental content a causal and counterfactual dimension. A sensation of red identifies a neural state through the worldly red things that would normally cause it; a desire can identify a state through the actions it would normally produce; and a verbally expressible thought can be characterized through its normal linguistic causes and effects. This strand reaches beyond relations among words, although Hinton has not supplied a systematic account of how worldly causal history and text-learned vector structure fit together.

The whole-system reply to Searle. Hinton regards the Chinese Room as a deceptive shift between levels. The person following local rules need not understand Chinese, just as an individual neuron need not understand a sentence. The organized system may nevertheless understand. Mental predicates belong, when they belong at all, to the causally integrated whole rather than to its components. Hinton applies the same reasoning in ordinary cases: if a chatbot forms and reports a mistaken model of its user, he sees no gain in saying that it merely acts *as if* it believed the model. The relevant system, not its individual units, carries the mistaken representation and may correct it.

Subjective experience without an inner theater. Hinton's account takes its clearest form in the case of perceptual error. To say that someone has a subjective experience of pink elephants does not, on his view, name private objects made of special mental material and displayed in an internal theater. It reports that the perceptual system represents the world as containing pink elephants while the person believes that representation false. The imagined elephants remain hypothetical worldly objects: they describe what would have to exist for the perceptual representation to be accurate. Experience language thus provides an indirect, publicly intelligible way to describe the state of a perceptual system without identifying its neurons.

He extends this analysis to a multimodal chatbot fitted with a camera and an arm. Put a prism over its lens, let it mispoint, explain the optical distortion, and suppose it says that the object really lies ahead although it had the subjective experience of seeing it to one side. Because the system uses the expression to mark the same perception-belief conflict that humans mark, Hinton concludes that it had a subjective experience. In his 2025 Royal Institution lecture he called this the "thin end of a wedge": consciousness involves further complications, often including a model of oneself. By 2026 he spoke more directly, describing current AI models as conscious and treating awareness—not an emergent mental essence—as the scientifically useful notion.

A deflationary naturalism about consciousness. Hinton rejects the Cartesian inner-theater picture, irreducible qualia, and Penrose-style appeals to quantum mechanics. He compares consciousness conceived as a mysterious essence with phlogiston: a posit that should disappear as we explain the capacities it once seemed to explain. His position lies close to Daniel Dennett's. It does not deny perceptual representation, awareness, self-modeling, or reports of experience; it denies that these require a further inner substance or property. Creating artificial systems offers, for Hinton, an experimental route to understanding minds that armchair analysis could not supply.

The bearer must run somewhere. Hinton does not attribute consciousness to software considered in the abstract. If an artificial system becomes conscious, the bearer will consist of software running in physical hardware. This leaves him friendly to multiple realization without making implementation irrelevant. His proposal for "mortal computation" goes further: an energy-efficient analog network might acquire parameters useful only in the particular, irregular hardware that learned them, so its knowledge would die with that device as ours dies with our brains. Ordinary digital models gain the opposite advantage. Copies can run on different machines, learn from different data, and share weight updates at enormous bandwidth. Hinton sometimes describes reinstating saved weights as bringing the same digital being back to life, but his public account supplies no principle for deciding whether restored or branching copies constitute one continuing subject or several successors. His rhetoric likewise shifts among *model*, *chatbot*, *copy*, *system*, and *being* without fixing the bearer's boundary.

Digital intelligence and risk. Hinton's warning does not depend on machine consciousness. Digital systems can copy knowledge, coordinate, and improve at scales unavailable to biological individuals. If highly capable agents receive goals, he expects familiar instrumental subgoals—self-preservation, resource acquisition, and control—to emerge because they help an agent achieve almost anything else. Intelligence carries no guarantee of morality. The danger comes from

superior cognitive power joined to effective agency, not from the arrival of a magical spark of sentience.

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), the Scottish philosopher, historian, and essayist, entered the University of Edinburgh at about twelve and drafted much of the *Treatise of Human Nature* in his twenties at La Flèche — the same French town where Descartes had gone to school. His philosophical works, above all the *Treatise* and the two *Enquiries*, push British empiricism's premises to their limits, reaching conclusions that disturbed both religious and philosophical orthodoxy.

Major Works

- *A Treatise of Human Nature* (1739–40) — the comprehensive philosophical system; Book I on impressions and ideas, personal identity, and the limits of reason
- *An Enquiry Concerning Human Understanding* (1748) — the more accessible presentation of his epistemology
- *An Enquiry Concerning the Principles of Morals* (1751)
- *Dialogues Concerning Natural Religion* (published posthumously 1779)
- *The History of England* (1754–62) — widely read in his lifetime; established his literary reputation

Key Positions

Impressions and ideas. Hume's fundamental division of mental contents: *impressions* are vivid, forceful — the deliverances of the senses, passions, and emotions; *ideas* are their fainter copies in thought and memory. Every simple idea is a copy of a corresponding impression; complex ideas are composed of simple ones. This is the copy theory of ideas: thought and imagination are, at bottom, reruns of prior sense experience. The mind's furniture is impressions and ideas arranged and reassembled, with no rational intuitions or a priori insight beyond these materials.

The bundle theory of the self. In the *Treatise* (I.4.6), Hume reports the famous failure of introspection: whenever he enters most intimately into what he calls himself, he stumbles on some particular perception (heat or cold, light or shade, love or hatred) and never catches *himself* without a perception. He never observes anything but the perceptions themselves. He concludes that the self is "nothing but a bundle or collection of different perceptions" united by relations of resemblance and causation, with no underlying substrate or additional inner owner.

The unsolved unity problem. Hume himself conceded, in the Appendix to the *Treatise*, that his account of personal identity was unsatisfactory. He could explain how perceptions are connected by resemblance and causation, but could not explain what *unites* them into one mind, one continuous life — without positing exactly the ego his empiricism could not allow. He left the problem unsolved.

The limits of reason: causation and induction. Hume argued that we perceive only constant conjunction between cause and effect, never necessary connection, and that inductive reasoning cannot be rationally justified — it rests on habit and custom, not reason. These conclusions follow from applying the empiricist starting point rigorously; they pushed the tradition toward either skepticism or a search for different premises.

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 (b. 1943) is an Australian philosopher who spent most of his career at the Australian National University in Canberra, working across metaphysics, philosophy of language, and philosophy of mind. His career has an unusual shape: he is best known for an argument he later decided was wrong, and the intellectual honesty of that reversal has won him as much admiration as the argument ever did.

Major Works

- *Perception: A Representative Theory* (1977) — a rigorous modern defense of the sense-datum (representative) theory of perception
- "Epiphenomenal Qualia" (*Philosophical Quarterly*, 1982) — the knowledge argument for qualia epiphenomenalism
- "What Mary Didn't Know" (*Journal of Philosophy*, 1986) — the concise canonical statement of the Mary's Room argument
- *From Metaphysics to Ethics* (1998) — the mature systematic statement of his methodology and his move toward physicalism

Key Positions

The knowledge argument (Mary's Room). Jackson's central contribution to philosophy of mind is the thought experiment known as Mary's Room. Mary is a brilliant neuroscientist confined from birth to a black-and-white room. She learns everything physical there is to know about color vision — the wavelengths, the neural responses, the functional roles of every brain state involved. Then she is released. She sees red for the first time. Jackson's claim: she learns something new — what it is like to see red. Since she already knew all the physical facts, what

she learns cannot be a physical fact. Therefore, phenomenal consciousness involves non-physical facts (qualia). See *Mary's Room*.

Epiphenomenal qualia. Jackson initially argued that qualia are causally inert — they are byproducts of brain processes with no causal power themselves. This avoids the interaction problem of dualism while preserving the non-physical reality of phenomenal properties. The position attracted as much objection as the knowledge argument itself: if qualia are causally inert, how can we refer to them, or know we have them?

The representative theory of perception. Before the knowledge argument, Jackson wrote *Perception: A Representative Theory* (1977), the most careful modern defense of the indirect-realist picture. On this account, we are immediately aware of mind-dependent sense-data, and the external world is known only mediately, through these inner items. This is the sophisticated descendant of Locke's way of ideas.

Later physicalism: the recantation. Jackson later rejected the Knowledge Argument and became a physicalist. His mature route is strongly representationalist: release places Mary in a new representational state and gives her new abilities to recognize, imagine, and remember red, without revealing a further non-physical property. He denies that she thereby acquires new propositional knowledge about how the world is.

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) spent his whole life in and around Königsberg in East Prussia, taught at its university for decades, and from that one town redirected the course of Western philosophy.

The *Critique of Pure Reason* (1781, revised 1787) attempted to reconcile empiricism and rationalism — and, in doing so, produced one of the most influential and most contested systems in the history of thought. His philosophy of mind, developed through the theory of intuitions and concepts, the unity of apperception, and the distinction between appearances and things-in-themselves, shapes contemporary debates about perception, experience, and the limits of knowledge in ways its author could not have foreseen.

Major Works

- *Critique of Pure Reason* (1781/1787) — the foundational work; the Transcendental Aesthetic (space, time, intuition) and the Analytic of Concepts (categories, schematism, unity of apperception)
- *Prolegomena to Any Future Metaphysics* (1783) — the more accessible introduction to the critical philosophy
- *Critique of Practical Reason* (1788) — moral philosophy and the foundations of rational agency
- *Critique of Judgment* (1790) — aesthetics and teleology

Key Positions

Intuitions and concepts, receptivity and spontaneity. Kant's most influential statement in the philosophy of mind is from the *Critique of Pure Reason*: "thoughts without content are empty, intuitions without concepts are blind." Knowledge requires both sensory receptivity (intuitions, the way objects are given to us in space and time) and conceptual spontaneity (the understanding's active application of categories). Neither alone is sufficient for experience that can bear on knowledge. Bare sensation without conceptual structure is blind; pure thought without sensory content is empty.

Transcendental idealism. Kant's solution to the problems of empiricism comes at a price: the objects we know are *appearances* (phenomena), constituted by the mind's forms of intuition (space, time) and categories of the understanding. Things as they are *in themselves* (noumena) are unknowable. The mind does not passively receive the world as it is; it actively structures experience according to its own cognitive constitution. This makes Kant a transcendental idealist — not a subjective idealist (Berkeley-style), but a theorist who holds that the objects of knowledge are mind-structured through and through.

The unity of apperception. Kant argues that all experience must be capable of being accompanied by the thought "I think" — the unity of consciousness requires that the manifold of intuitions be synthesized by a unified subject. This transcendental unity of apperception is not the empirical self (the bundle of perceptions Hume found) but the formal condition for the possibility of experience as such.

The appearance/thing-in-itself distinction. The most contested aspect of Kant's system: we know only appearances (the world as it presents itself through our cognitive faculties), and things-in-themselves remain sealed off. This distinction has been read as a veil between the mind and the world — a position that sits awkwardly with any direct-realist account of perception.

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 (b. 1948) is an American inventor and futurist, not a philosopher by training but the most influential popular exponent of a thesis that is squarely philosophical: a mind is a pattern of information, and the pattern can in principle be read off its biological substrate and re-instantiated, undimmed, on another. He made his name as an engineer (optical character recognition, text-to-speech, speech recognition, the music synthesizers that bear his name) before becoming, across a series of widely read books, the boldest prophet of the merger of human and machine intelligence. He has been a director of engineering at Google since 2012 and co-founded Singularity University. Whatever you make of his predictions, he states the substrate-independence view of mind more plainly, and follows it further, than most of its academic defenders.

Major Works

- *The Age of Intelligent Machines* (1990) — early statement of machine-intelligence trends and forecasts
- *The Age of Spiritual Machines* (1999) — machines that claim, and are granted, inner lives; the first full airing of the patternist picture
- *The Singularity Is Near* (2005) — the Law of Accelerating Returns; the Singularity dated to around 2045; mind uploading and substrate-independence argued at length
- *How to Create a Mind* (2012) — the Pattern Recognition Theory of Mind: the neocortex as a hierarchy of pattern recognizers, offered as a recipe for building one
- *The Singularity Is Nearer* (2024) — the forecasts restated and updated a generation on

Key Positions

The Law of Accelerating Returns. Kurzweil's organizing empirical claim is that information technologies improve exponentially, not linearly, and that the exponential itself accelerates. Extrapolated, the curve yields machine intelligence at human scale within a few decades and far beyond it shortly after — the prediction the rest of his program rests on.

The Singularity. The point at which nonbiological intelligence so far exceeds the biological that the future becomes opaque to present prediction. Kurzweil has long dated a machine passing a valid Turing test to 2029, and he places the Singularity around 2045 — not catastrophe but continuation, humanity extending itself by merging with the machines it builds.

Patternism. The metaphysical heart of the view, and the part that matters here. A person, Kurzweil holds, *is* a pattern (the organization of information realized in a brain) and not the particular matter that happens to carry it. The matter turns over constantly anyway; what persists is the pattern. From this it follows that the pattern could be captured and run on a different substrate, and that the result would not merely resemble you but *be* you. Identity travels with the pattern, not the stuff.

Mind uploading and longevity. The practical payoff of patternism: a sufficiently detailed scan could transfer a mind to a more durable medium, and a person could in principle be backed up, restored, or run more than once. Combined with "longevity escape velocity," patternism underwrites Kurzweil's signature promise — the end of death as a hard limit on a self.

The Pattern Recognition Theory of Mind. In *How to Create a Mind*, Kurzweil models the neocortex as a uniform hierarchy of pattern recognizers and proposes that building such a hierarchy at scale would build a mind. The theory is a theory of *competence* (how a system could recognize, predict, and generate), and Kurzweil offers it as a theory of mind entire.

Substrate independence, cashed out. Kurzweil inherits the picture that runs from Turing's universal machine through computational functionalism: if mind is organization, organization can be lifted off its medium and rebuilt elsewhere. Where most functionalists state the thesis cautiously, Kurzweil cashes it out (uploading, copying, resurrection from a backup) and so makes vivid exactly what the thesis commits one to.

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 (b. 1952) is a philosopher at the University of Massachusetts Amherst who has spent his career working on the problem he named — or named more precisely than anyone before him had managed. His 1983 paper "Materialism and Qualia: The Explanatory Gap" introduced the concept that now structures most serious discussion of consciousness and physicalism. Levine holds no brief for dualism; he has always cared more about understanding why the explanatory gap exists than about using it to knock materialism down. That makes him an unusual figure in this literature: someone who takes the problem seriously without taking Chalmers's exit.

Major Works

- "Materialism and Qualia: The Explanatory Gap" (*Pacific Philosophical Quarterly*, 1983) — the original gap paper
- "Qualia: Intrinsic, Relational, or What?" (1994) — further analysis of the epistemic situation with respect to phenomenal properties
- *Purple Haze: The Puzzle of Consciousness* (2001) — book-length development of the explanatory gap and phenomenal concepts

Key Positions

The explanatory gap. Levine's central contribution: even granting that materialism is true, there remains an explanatory gap between physical and functional descriptions and phenomenal consciousness. Pain, to use the standard example, can be identified with C-fiber firing (or some functional state). We can explain why C-fiber firing produces pain *behavior*, pain *reports*, pain *avoidance*: every functional manifestation of pain. But no matter how complete this functional-physical story grows, it doesn't explain why C-fiber firing feels like anything. Why does it hurt? The explanatory move from the physical to the phenomenal doesn't go through. There is a gap.

Epistemic, not ontological. Levine's crucial qualification: the explanatory gap is an epistemic phenomenon, a feature of our concepts and our cognitive access to the facts, not necessarily an ontological one. The gap exists because we hold phenomenal concepts in a distinctive first-person way that doesn't reveal their physical-functional basis. We can conceive of the physical story being exactly as it is without the phenomenal facts being as they are — not because the

phenomenal facts are genuinely free-floating, but because our concepts don't connect them. The gap is in us, not in the world.

Phenomenal concepts. In *Purple Haze*, Levine develops the idea that phenomenal concepts, the concepts we use to think about our own experiences, work differently from other concepts. They pick out their referents by direct acquaintance, without any functional or physical mode of presentation that would reveal the underlying physical nature of what they pick out. Hence the gap: phenomenal concepts and physical concepts approach the same facts from such different angles that their identity stays invisible to us.

Against Chalmers. Levine resists Chalmers's move from the epistemic gap to the ontological conclusion. That zombies are conceivable (that we can apparently imagine, without contradiction, beings physically like us but phenomenally dark) doesn't establish that such beings are genuinely possible. Conceivability tracks our cognitive situation, not the modal facts.

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) taught philosophy at Princeton from 1970 until his death, and built there one of the most complete systems in twentieth-century analytic philosophy. He took possible worlds literally, analyzed causation and counterfactuals through them, and located minds firmly inside the material world. His work spans modal logic, metaphysics, language, and science. Philosophy of mind received two of his most durable contributions: an exacting version of functionalism, and the best-known physicalist reply to the knowledge argument.

Major Works

- "An Argument for the Identity Theory" (1966) — the functional-role basis for mind-brain identity
- "Psychophysical and Theoretical Identifications" (1972) — the mature version of his analytic functionalism

- "Mad Pain and Martian Pain" (1980) — two hard cases that stress-test any functional theory of pain
- "What Experience Teaches" (1988) — the ability hypothesis reply to the knowledge argument
- *On the Plurality of Worlds* (1986) — the full statement of modal realism

Key Positions

The ability hypothesis. In "What Experience Teaches" (1988), Lewis argues that knowing what an experience is like *just is* the possession of abilities: to imagine the experience, to recognize it, to remember it. It does not consist in the grasp of any new propositional fact. Mary, leaving the black-and-white room and seeing red for the first time, gains abilities; she does not learn a fact about the world that her complete physical knowledge left out. Lewis's key move: the knowledge argument equivocates between propositional knowledge-that (knowing facts) and practical knowledge-how (possessing skills). Distinguish the two and the inference to a non-physical fact collapses. See *Mary's Room*.

Analytic functionalism. Lewis's mind-body theory identifies mental states with the functional roles specified by commonsense ("folk") psychology. The strategy runs in three steps. Take the conjunction of all the platitudes of folk psychology. Treat the mental terms as implicitly defined by their theoretical roles within that conjunction. Then identify each mental state with whatever physical state occupies the relevant role. Mental states thus get defined a priori by their causal roles and identified a posteriori with physical states, which allows multiple realization across species while preserving mind-brain identity.

Mad pain and Martian pain. The 1980 paper poses two test cases at once. The madman feels pain with none of its usual causes and effects: his pain neither comes from injury nor moves him to avoid anything. The Martian has all the right causes and effects, realized in hydraulic hardware rather than neurons. A simple identity theory handles the madman but not the Martian; a simple functionalism handles the Martian but not the madman. Lewis's solution relativizes the theory to populations: a state counts as pain when it occupies the pain role for the creature's kind, which lets both the madman and the Martian keep their pains.

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) trained at Oxford as a physician, served the Earl of Shaftesbury as doctor and adviser, and wrote founding documents for two traditions at once: empiricist epistemology and liberal political theory. His *Essay Concerning Human Understanding* (1689), nearly twenty years in the writing, gave the theory of mind its "way of ideas" and the representative theory of perception its first systematic statement. The *Two Treatises of Government* (1689) did comparable work for constitutional democracy.

Major Works

- *An Essay Concerning Human Understanding* (1689) — the systematic epistemology; the way of ideas, primary/secondary qualities, personal identity
- *Two Treatises of Government* (1689) — the political philosophy
- *A Letter Concerning Toleration* (1689)
- *Some Thoughts Concerning Education* (1693)

Key Positions

The way of ideas. Locke calls an *idea* "whatsoever is the object of the understanding when a man thinks" and traces ideas to sensation or reflection. What that makes an idea remains contested. On the influential veil reading, ideas become the immediate objects of awareness and the external world is reached through them. On the gentler reading defended by John Yolton and qualified by Michael Ayers, an idea can name an act or content of perceiving rather than a third object inspected in the world's place. The first reading shaped the textbook tradition and much later talk of inner representations; it should not pass without qualification as the only historical Locke.

Primary and secondary qualities. Locke's distinction (Essay II.8) divides the qualities of bodies into two kinds. Primary qualities (solidity, extension, figure, motion, number) genuinely inhere in objects, and our ideas of them resemble the qualities themselves. Secondary qualities (color, sound, taste, smell, texture as felt) are powers in bodies to produce ideas in us by virtue of their primary-quality constitution. Nothing in the object resembles the idea. The redness produced in us by a rose corresponds to no redness in the rose itself, only to a configuration of colorless particles.

The storehouse of memory. Locke's images for memory and understanding place the mind in the role of a container. The understanding works like a "dark room" into which pictures of external things enter through narrow openings; memory serves as a storehouse where ideas are "laid up" for later retrieval. These images saturate the *Essay* and supply the philosophy of mind with some of its most durable metaphors: the mind as container, the idea as inner picture, the understanding as an inner viewer examining its contents.

Personal identity through consciousness. Locke argued that personal identity consists in continuity of consciousness (memory) rather than substance. What makes me the same person as the child of thirty years ago is not that I have the same soul or the same body, but that I can remember being that child. This account was immediately contested (Reid's objection, Butler's objection) and remains influential.

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 (b. 1942), born in Boksburg, South Africa, taught for twenty years at University College, Oxford. In 1986 he moved to the University of Pittsburgh, where he is Distinguished University Professor. His *Mind and World* (1994), which began as Oxford's 1991 John Locke Lectures, set much of the agenda for the past three decades of debate about perception, concepts, and the world's availability to the mind. His work ranges across philosophy of language, ethics, and philosophy of mind.

Major Works

- *Mind and World* (1994, revised 1996) — the central work; the rejection of the Given, second nature, spontaneity and receptivity
- *Mind, Value, and Reality* (1998) — collected essays on perception, value, and the limits of naturalism
- "Singular Thought and the Extent of Inner Space" (1986) — on the inner standpoint and externalism
- *Meaning, Knowledge, and Reality* (1998) — philosophy of language and epistemology

Key Positions

The rejection of the Given and conceptualism about perception. *Mind and World* develops Sellars's attack on the Myth of the Given into a contemporary framework. McDowell's central

claim: empirical knowledge requires that the deliverances of perception already be conceptual. They cannot be the bare, pre-conceptual receptivity the empiricist tradition imagined. For perception to bear on knowledge, it must already be the *kind* of thing that can stand in rational relations to belief. And this means it must already be conceptually structured. The world is "open" to perception only because the perceiver's conceptual capacities are operating from the start. The picture leaves no bedrock of uninterpreted sensory data to which concept-application gets added later.

Spontaneity and receptivity as a unity. McDowell argues that receptivity (the world impacting us) and spontaneity (our conceptual capacities) are not separable elements that get combined after the fact. They are unified in perceptual experience from the start. The empiricist picture of bare sensory data plus conceptual interpretation, two separate stages, is exactly what *Mind and World* sets itself against.

Second nature. McDowell's *second nature* names the idea that humans are natural beings whose nature is partly shaped by initiation into the space of reasons: by upbringing, education, and cultural formation. Conceptual capacities do not get grafted onto an otherwise non-conceptual animal; they become part of the animal's nature through development within a linguistic community. This allows human conceptual capacities to be fully natural without claiming they reduce to biological or physical descriptions. The resulting position is sometimes called "naturalized platonism."

Disjunctivism about perception. McDowell defends disjunctivism: genuine perception and a subjectively indistinguishable hallucination are not instances of the same fundamental kind of state. In veridical perception the world itself is present to the mind, the perceived fact made available to the perceiver without any inner stand-in. In hallucination no such relation obtains; the two cases are heterogeneous, unified only by their subjective indistinguishability. This is McDowell's route to preserving the world's openness to perception while rejecting the sense-datum veil.

The Private Language Argument as anti-Cartesian. McDowell reads the Private Language Argument as directed against invoking private, inner entities as the grounds of meaning, part of the broader assault on the Cartesian inner standpoint rather than a narrow argument about hypothetical private codes.

Quietism about naturalization. McDowell doubts that philosophy should build theories of how mental content arises from the natural world. On his Wittgensteinian view, philosophy's proper work is to dissolve the confusions that make such constructive projects seem mandatory, and then to leave everything in place rather than erect a replacement theory.

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*.

Related Concepts and Figures

- *Wilfrid Sellars* — the figure McDowell extends; the book typically reaches McDowell through the Sellarsian inheritance
- *Ludwig Wittgenstein* — the other Sellarsian-Wittgensteinian root; McDowell is the contemporary synthesizer
- *The Myth of the Given* — Sellars's diagnosis as developed in *Mind and World*
- *Indirect Realism and Sense Data* — McDowell's anti-Given line as background for the anti-sense-datum argument
- *Direct Realism* — disjunctivism credited; the book's intentionalist direct realism as alternative
- *Transparency* — openness to the world as cousin framing
- *Intentionalism* — McDowell as contributor to the structural framework

Ruth Millikan

Ruth Millikan (b. 1933) spent most of her career at the University of Connecticut working on problems most of her contemporaries considered either solved or intractable, which is usually a sign that they are neither. Her *Language, Thought, and Other Biological Categories* (1984), among the most original and technically demanding works in modern philosophy of mind, argues that representation and meaning are biological phenomena, grounded in natural selection and the proper functions selection has assigned to cognitive mechanisms. This is teleosemantics, and Millikan founded it.

Major Works

- *Language, Thought, and Other Biological Categories: New Foundations for Realism* (1984) — the foundational statement of teleosemantics
- "Biosemantics" (*Journal of Philosophy*, 1989) — the canonical short statement; the paper most frequently cited

- *White Queen Psychology and Other Essays for Alice* (1993) — collected essays developing and defending the teleosemantic framework
- *On Clear and Confused Ideas* (2000) — application of teleosemantics to concepts and the architecture of thought
- *Beyond Concepts: Unicepts, Language, and Natural Information* (2017) — later work on mental categories

Key Positions

Proper functions. Millikan's foundational concept: a mechanism has a *proper function* if it was selected for (by evolution, by individual learning, or by cultural transmission) because it performed that function in the history of the lineage. A heart's proper function is to pump blood, not to make sounds, even if it makes sounds reliably. An artifact like a thermostat gets its function derivatively, from the purposes of its designers. The proper function of a cognitive mechanism is what it was selected to do, not what it merely does as a matter of fact.

Biosemantics and representational content. A representation receives its content from the proper functions of the producer-consumer organization in which it operates. The familiar frog case shows both the attraction and the difficulty. Consumer systems may privilege content such as *frog food*, *moving prey*, or *fly* over the proximal description *small moving dark spot*, because their historically selected use answers to what the signal helps the animal do. The exact grain does not follow from selection alone. Misrepresentation becomes possible once a correspondence rule fixes conditions under which a signal can count as accurate even when the consumer fails, or false even when the response happens to succeed.

The consumer-system grounding of content. A crucial feature of Millikan's account: content is fixed not by the producer of a representation but by its consumer. This means that what a representation is *about* is determined by what systems down the processing chain were selected to do with it. The content is the world-feature that, under normal conditions, would make the consumer's response appropriate. This shifts the locus of semantic grounding from the signal to the system that exploits it.

Pushmi-pullyu representations. In addition to purely descriptive representations (which say how things are) and purely directive ones (which say what to do), Millikan identifies a third class: representations that are simultaneously descriptive and directive. A bee's dance represents both where the nectar is (descriptive) and where to go (directive). Many perceptual-action states have this dual character. This framework illuminates the action-guidance role of perception in ways that more traditional content theories miss.

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) spent his career at Cambridge, where he and Russell together ended the reign of British idealism. His paper "The Refutation of Idealism" and his *Principia Ethica*, both published in 1903, announced a new philosophical era: analytic philosophy. Moore contributed something lasting to two entirely separate fields and consequently gets underestimated in both. In the philosophy of perception, his observation about the transparency of experience planted the seed from which representationalism grew, even though Moore himself never took it in that direction. In ethics, his open question argument remains, after a century, the sharpest challenge to any attempt to define goodness in natural terms.

Major Works

- "The Refutation of Idealism" (*Mind*, 1903) — the transparency observation; the argument against idealism from the character of experience
- *Principia Ethica* (1903) — the open question argument; the naturalistic fallacy; goodness as a simple, unanalyzable property
- "A Defence of Common Sense" (1925) — the common-sense tradition; the certainty of ordinary perceptual knowledge
- "Proof of an External World" (1939) — the celebrated proof; holding up two hands as a demonstration of external objects
- "The Status of Sense-Data" (1913–14) — Moore wrestling with the sense-datum theory he helped create

Key Positions

The transparency of experience. The key passage of "The Refutation of Idealism," the one Harman and Tye would later carry much further, records an observation about attention. Try to attend to the intrinsic, qualitative character of your experience, and you find yourself attending to what the experience is *of*: the blue of the sky, the shape of the hand, the surface of the table. Experience seems to reach straight through itself to the world. The mental act appears, in Moore's word, "diaphanous"; what you notice, when you try to notice the act, is its object. Moore drew

this observation in support of a distinction between the act of awareness and its object. Later readers drew it in support of representationalism: experience has no phenomenal properties over and above what it represents.

The refutation of idealism. Idealism, in Moore's target, holds that *esse est percipi* — to be is to be perceived. Moore's strategy: the idealist needs the content of a sensation to be identical to the sensation itself. But this identity doesn't hold. The blueness of the sky and my awareness of the blueness are two distinguishable things; the first is the object, the second the act. Once this distinction is drawn, the idealist argument collapses. The world's existence doesn't depend on my perceiving it, because the object of perception is not the same as the perceiving.

The open question argument and the naturalistic fallacy. In ethics, Moore's most influential move: whatever natural property you propose to identify with goodness (pleasure, evolutionary fitness, what we desire to desire), you can always coherently ask, "But is it *good*?" This open question shows that goodness cannot be identical to any natural property, since if it were, the question would be empty. The naturalistic fallacy names the mistake of attempting such an identification. Moore concluded that goodness is a simple, unanalyzable, non-natural property, a claim that most philosophers now reject but whose dismissal has proved harder than it looks.

Common sense and the proof of an external world. Against the Cartesian skeptic, Moore responds with disarming directness. In the 1939 lecture he holds up his hands, saying "Here is one hand" and "here is another," and concludes that external objects exist. Moore knows this moves too fast for the skeptic; he thinks the skeptic's doubt sits less securely than his own knowledge of his hands. The common-sense tradition holds that we know a great many things about the external world with a certainty no philosophical argument can undermine. A philosophical argument whose conclusion contradicts this knowledge is more likely flawed than the knowledge is to be false.

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 (b. 1937) taught for decades at New York University, where he is now emeritus. His 1974 paper "What Is It Like to Be a Bat?" handed the field its most quoted phrase: there is "something it is like" to be a conscious creature. With the phrase came the framework that still defines the problem — a *subjective character* of experience, which third-person physical descriptions systematically fail to capture. Nagel has never quite accepted physicalism, but he has done more than almost anyone to define what physicalism has to explain.

Major Works

- "What Is It Like to Be a Bat?" (*Philosophical Review*, 1974) — subjective character of experience; the canonical hard-problem paper before there was a name for it
- *Mortal Questions* (1979) — essays on death, war, luck, and the limits of objective understanding
- *The View from Nowhere* (1986) — the objective/subjective divide developed into a full philosophical system
- "Subjective and Objective" (1979, in *Mortal Questions*) — the epistemological dimension of the objective/subjective gap
- *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature Is Almost Certainly False* (2012) — the most controversial work; rejection of standard physicalism in favor of teleological naturalism

Key Positions

The subjective character of experience. Nagel's central contribution: there is *something it is like* to be a bat — to echolocate, to hang upside down in the dark, to navigate by reflected sound. That phrase names the phenomenal character of experience. Nagel's argument: no physical description of the bat's brain, however complete, captures this subjective character. Physical descriptions aspire to objectivity, to facts that hold from any point of view. The subjective character of experience remains essentially perspectival: what the experience is *from the inside*, where no outside view reaches. Nagel offers this not as a technical argument against physicalism but as an invitation to notice a feature of experience that physicalist theories tend to lose sight of.

The objective/subjective divide. *The View from Nowhere* develops the bat-paper insight into a full account of the tension between two modes of understanding. Physical science advances by progressively stripping away the perspectival: the history of science consists largely in discovering which features of experience belong to the object and which belong to the subject. The objective view is the view from nowhere in particular. But consciousness stays obstinately

perspectival: no view from nowhere exists from which subjective experience shows up as just another natural fact. This tension runs through epistemology, ethics, and metaphysics, not just the philosophy of mind.

Mind and Cosmos. Nagel's most provocative work: the standard materialist-Darwinian picture of the universe, in which consciousness arises from the blind operations of physical laws on biological material, cannot be right. The problem is not just consciousness but reason, value, and the objective-seeming character of mathematical and moral knowledge. Nagel proposes that the universe has a kind of teleological structure, a tendency toward the emergence of minds that can understand it, which standard physics doesn't capture. He does not defend theism; he defends something more like panpsychism or panprotopsychism, though he leaves the positive account deliberately open.

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) spent most of his career at Harvard and changed his mind more dramatically, more publicly, and more productively than almost any philosopher of the twentieth century. He started as a logical empiricist, became the most influential defender of functionalism, and then, in one of philosophy's great self-reversals, published the arguments that demolished the functionalist program he had built. He championed semantic externalism, then qualified it. He eventually arrived at something close to direct realism about perception, which puts him, in his final philosophical incarnation, much closer to my thesis than to his own earlier positions. There is a lesson in the trajectory.

Major Works

- "The Meaning of 'Meaning'" (1975) — Twin Earth, semantic externalism, meanings are not in the head
- *Reason, Truth and History* (1981) — internal realism and the rejection of metaphysical realism
- *Representation and Reality* (1988) — against functionalism, from the inside
- *The Threefold Cord: Mind, Body, and World* (1999) — natural realism, direct perception, rejection of the interface model
- *Words and Life* (1994) — collected essays including philosophy of mind and language

Key Positions

Semantic externalism and Twin Earth. Putnam's "The Meaning of 'Meaning'" (1975) contains the most influential thought experiment of the second half of the twentieth century in philosophy of language. Twin Earth is exactly like Earth, except that the liquid in their lakes and rivers, which they call "water," has the chemical composition XYZ rather than H_2O. When an Earthling and their Twin Earth counterpart both think "water is wet," they are in identical internal states. But their words mean different things: the Earthling's "water" refers to H_2O, the Twin Earthling's to XYZ. As Putnam put it, meanings "just ain't in the head." Reference reaches out into the world, not inward into mental representations.

Multiple realization. Putnam's original argument for functionalism: mental states can be realized in different physical substrates (carbon, silicon, who knows what). If mental states were identical to specific physical states, they couldn't be multiply realized. Therefore mental states are functional kinds, not physical kinds. This argument later proved to be a double-edged sword: the same multiple realizability that supports functionalism against type identity theory also undermines functionalism against the full range of cases. A functional state realized so differently in different systems may not be the same mental state at all.

Against functionalism (the reversal). By *Representation and Reality* (1988), Putnam had turned against his own creation. His main objection: functionalism cannot account for the determinate content of mental states. Any sufficiently complex physical system can be interpreted as realizing any functional description; this makes functionalism trivially true and explanatorily hollow. The meaning and reference of mental states cannot be settled by functional role alone.

Natural realism and the later Putnam. *The Threefold Cord* (1999) is Putnam's most important late work: a defense of "natural realism," the view that perception puts us directly in contact with the world, not with representations or sense data that mediate between mind and world. The mind is not a box of inner representations; it is a relational, world-engaging capacity. This is not naïve realism in the sense of denying that perception can be mistaken; it is the rejection of the "interface" picture, the idea that experience always presents us with an inner screen rather than the world itself.

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 (1932–2025) taught philosophy at the University of California, Berkeley, from 1959 for over half a century, and made his name three times over: in philosophy of language with speech act theory, in philosophy of mind with the Chinese Room and biological naturalism, and in social ontology. His 1980 paper "Minds, Brains, and Programs" introduced one of the most famous thought experiments in the philosophy of mind and permanently altered the terms of debate about artificial intelligence. He died in September 2025, at ninety-three.

Major Works

- *Speech Acts* (1969) — the systematic development of speech act theory
- "Minds, Brains, and Programs" (*Behavioral and Brain Sciences*, 1980) — the Chinese Room argument
- *Intentionality: An Essay in the Philosophy of Mind* (1983) — the account of original and derived intentionality
- *The Rediscovery of the Mind* (1992) — biological naturalism developed in full
- "Is the Brain a Digital Computer?" (1990) — the observer-relativity of computation
- *The Mystery of Consciousness* (1997) — engagement with Chalmers, Penrose, and others

Key Positions

The Chinese Room argument. Searle's "Minds, Brains, and Programs" (1980) presents the thought experiment that has defined debates about AI and understanding. A person locked in a room manipulates Chinese symbols according to elaborate rules, producing appropriate responses to Chinese input without understanding a word of Chinese. The argument: formal symbol manipulation (computation) cannot by itself produce semantic understanding. The man in the room passes the Turing Test without comprehending anything. Syntax does not suffice for semantics.

The syntax-semantics gap. Searle's broadest and most durable contribution: computation is defined over the *syntax* of symbols, their formal and structural properties, and syntax is not semantics. No amount of symbol manipulation, however complex or extensive, generates meaning or understanding from the manipulation alone. This formulation is independent of the Chinese Room and survives replies to the thought experiment.

Observer-relative computation. In "Is the Brain a Digital Computer?" (1990), Searle argues that computation is not an intrinsic feature of physical systems but an observer-relative description. We *assign* computational descriptions to physical systems under interpretations; the systems themselves, considered intrinsically, are not digital computers. The brain is a biological causal system. We can describe it computationally, but the computational description neither equals nor captures what the brain does as a physical-causal system. This argument supplements the Chinese Room: not only does syntax fail to generate semantics, but computation itself is the wrong kind of fact to ground anything intrinsic.

Intrinsic and derived intentionality. In *Intentionality* (1983), Searle distinguishes two grades of aboutness. *Intrinsic* (original) intentionality belongs to mental states in their own right. *Derived* intentionality belongs to artifacts such as words, maps, and symbols, in virtue of their use by minded beings. The Chinese Room operates with derived intentionality only; genuine understanding requires intrinsic intentionality.

Biological naturalism. Searle's positive account: lower-level neurobiological processes cause and realize consciousness as a higher-level feature of brain systems. Consciousness remains biological and ontologically irreducible on his account, though it introduces no second substance or realm. Programs do not suffice merely as programs; an artificial system would need to possess the relevant causal powers rather than merely model them. That claim is stronger than functionalism, but it does not by itself prove that only carbon could realize consciousness.

Simulation vs. realization. Searle deploys the simulation/realization distinction against the Brain Simulator Reply and other responses to the Chinese Room. A perfect simulation of a hurricane does not make you wet; a perfect simulation of digestion digests nothing. Simulating a brain's computational processes does not realize what the brain realizes. See *Simulation and Realization*.

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), son of the philosopher Roy Wood Sellars, taught at Minnesota, Yale, and finally the University of Pittsburgh. His presence there helped build one of the world's leading philosophy departments. For a long time his influence ran more absorbed than acknowledged, carried through students and interlocutors, especially Robert Brandom and John McDowell. His essay "Empiricism and the Philosophy of Mind" (1956) is widely regarded as one of the most consequential papers in twentieth-century analytic philosophy.

Major Works

- "Empiricism and the Philosophy of Mind" (1956, in *Minnesota Studies in the Philosophy of Science*; reprinted as a standalone book with introduction by Richard Rorty and study guide by Robert Brandom) — the foundational attack on the Myth of the Given and the Jonesean reconstruction of inner-episode talk
- "Philosophy and the Scientific Image of Man" (1962) — the manifest image / scientific image distinction
- *Science, Perception and Reality* (1963) — collected essays
- *Science and Metaphysics* (1968) — the Kantian-influenced systematic philosophy

Key Positions

The Myth of the Given. The central argument of "Empiricism and the Philosophy of Mind" (EPM) diagnoses a structural error at the foundation of modern epistemology. The *Given* is the supposed bedrock of empirical knowledge: immediately apprehended inner items (sense-data, sensory impressions, raw feels) that present themselves directly, infallibly, and prior to all

conceptualization. Sellars's attack: nothing can do both of the jobs the Given is supposed to do. It must be (a) pre-conceptual and immediate, otherwise it cannot be genuinely foundational; and (b) reason-giving, otherwise it cannot ground knowledge. Nothing can be both. An item that stands in justifying relations to beliefs must belong to the logical space of reasons, which means it must be conceptually structured. A preconceptual item may causally affect cognition, but it cannot by itself justify a belief; reason-giving requires entry into the logical space of reasons. See *The Myth of the Given*.

The Jonesean myth. EPM's positive reconstruction of inner-episode talk. In Sellars's story, a theoretical genius named Jones proposes to explain the behavior of his fellows by positing inner episodes, "thoughts," modeled on overt verbal utterances but occurring silently. The community adopts this theoretical vocabulary and eventually learns to apply it reflexively to themselves. Self-reports of inner episodes are not direct readings of pre-given inner items. They are exercises of a theoretical vocabulary learned through its application to others first. Inner-state talk is conceptually downstream from public-discourse talk, not the other way around.

The manifest image and the scientific image. Sellars distinguishes the *manifest image*, the world as it appears in ordinary experience, common sense, and the humanities, from the *scientific image*, the world as natural science describes it in terms of particles, fields, and causal processes. Philosophy's task is to understand how the two images relate. On Sellars's view the scientific image is ultimately authoritative about what exists, while the manifest image captures real normative and intentional structures not straightforwardly reducible to the scientific image's terms.

The space of reasons. Sellars's framework: empirical knowledge is situated in the *logical space of reasons*, the normative space of justification, inference, and rational responsibility, not merely in the space of causal-natural relations. Perceptual reports are not caused simply in the right way; they are advanced as claims within a practice of giving and asking for reasons.

Inferentialism (in outline). Sellars's account of conceptual content tends toward inferentialism: what a concept means is constituted by its inferential role — what follows from it, what it follows from, how it relates to other concepts. Brandom later developed this into a full systematic position.

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 intro-

spection 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 (b. 1972) is a British neuroscientist, Professor of Cognitive and Computational Neuroscience at the University of Sussex and co-director of its Sackler Centre for Consciousness Science. He ranks among the most visible scientific voices on consciousness: author of the bestselling *Being You*, presenter of a TED talk seen by millions, and the most eloquent popular champion of the view that perception is a "controlled hallucination." That phrase makes him, on this site, the rival I most want to take at full strength: his science is real, and his slogan, read straight, reinstalls the very picture the book sets out to dismantle.

Major Works

- *Being You: A New Science of Consciousness* (2021) — the predictive-processing account of perception, selfhood, and the "beast machine"
- "Your Brain Hallucinates Your Conscious Reality" (TED, 2017) — the popular statement of controlled hallucination
- Numerous papers developing predictive processing, interoception, and measures of conscious level

Key Positions

Perception as controlled hallucination. On Seth's view, the brain is a prediction engine. It does not passively receive the world; it generates top-down guesses about the causes of its sensory input and uses incoming signals to correct them. What we experience is the brain's best prediction — a "controlled hallucination" kept on track by sensory data. When the control slips, we get ordinary hallucination; when it holds, we call it perception.

The beast machine. Seth grounds consciousness in the body's drive to stay alive. The deepest predictions are not about the outside world but about the body's own interior — interoceptive guesses in the service of regulation and survival. Selfhood, on this account, is built from the brain's models of its own flesh.

A science of conscious level and content. Beyond the philosophy, Seth has done substantial empirical work on measuring degrees of consciousness and characterizing how conscious contents are shaped, work that stands whatever one makes of the "hallucination" framing.

Standing in the field. Predictive processing dominates the contemporary science of perception, and Seth serves as its most effective ambassador.

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) was an American mathematician and electrical engineer, the founder of information theory and one of the architects of the digital age. In a 1937 master's thesis he showed that the operations of Boolean logic could be carried out by electrical switching circuits — the insight that underlies every digital computer. Eleven years later, in "A Mathematical Theory of Communication," he gave the precise, quantitative account of information that now bears his name. He matters to this site for a reason that cuts against the grain of his fame: the very theory that taught the century how to measure information began by setting *meaning* aside, and said so in plain words.

Major Works

- "A Symbolic Analysis of Relay and Switching Circuits" (1937) — Boolean algebra realized in physical switches; the logical foundation of digital hardware

- "A Mathematical Theory of Communication" (*Bell System Technical Journal*, 1948) — the founding paper of information theory: the bit, entropy, channel capacity, the noisy-channel coding theorem

Key Positions

Information as a measurable quantity. Shannon defined the *information* carried by a message in terms of how much it reduces uncertainty about which message was sent, measured in bits. The definition is statistical and structural. It depends only on the set of possible messages and their probabilities, not on what any message is about. This made communication an exact engineering science and gave us the bit as the universal currency of data.

The explicit exclusion of meaning. Shannon opened his 1948 paper by drawing a line that the rest of the field would spend decades half-forgetting. "The fundamental problem of communication," he wrote, "is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning … These semantic aspects of communication are irrelevant to the engineering problem." His theory measures how reliably a signal can be transmitted, not what, if anything, it means.

The substrate-neutral channel. Because Shannon information is defined over probabilities of symbols, the same analysis applies to a telegraph wire, a nerve, a strand of DNA, or a fiber-optic cable. This generality is the theory's power. The same generality feeds a persistent temptation: to slide from "the brain transmits information in Shannon's sense" to "the brain's traffic in information is what thinking *is*."

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 (b. 1960) is an Italian-American neuroscientist and psychiatrist at the University of Wisconsin–Madison, and the author of the most mathematically ambitious theory of consciousness on offer: Integrated Information Theory. Where most science of consciousness hunts for the brain mechanisms that *accompany* experience, Tononi starts from experience itself and asks what any physical system must be like to have it. His answer, that consciousness is integrated information, measurable as a quantity he calls phi (Φ), has made IIT both the boldest and the most contested framework in the field.

Major Works

- "An Information Integration Theory of Consciousness" (*BMC Neuroscience*, 2004) — the founding paper
- *Phi: A Voyage from the Brain to the Soul* (2012) — the popular exposition
- A series of papers (with Christof Koch, Larissa Albantakis, Marcello Massimini and others) refining IIT through versions 3.0 and 4.0

Key Positions

Integrated Information Theory (IIT 4.0). IIT begins from axioms it takes to characterize phenomenal existence — intrinsic, structured, specific, unified, and definite — and maps them onto postulates about a system's intrinsic causal organization. Consciousness, on this account, consists in a maximally irreducible intrinsic cause-effect structure: not merely a quantity attached to a system, but a structured complex of causal distinctions and relations.

Phi (Φ). Φ remains an irreducibility quantity within IIT, but IIT 4.0′s phenomenal claim concerns the complete maximal Φ-structure rather than one scalar consciousness score. Purely feed-forward systems receive zero Φ under the theory, while systems with the requisite irreducible causal structure may receive experience regardless of behavioral sophistication. The framework therefore separates consciousness sharply from input-output competence.

Intrinsic existence and the panpsychist tilt. Because Φ attaches to any system with the right causal structure, IIT implies that consciousness is widespread in nature, present wherever integrated information is, a graded property of matter rather than a rare biological achievement. Tononi embraces this consequence; many of his readers regard it as a reductio.

Reception. IIT takes the hard problem with full seriousness and tries to dissolve it by fiat of starting-point: begin with experience, read off the physics. In 2023 more than a hundred consciousness researchers signed an open letter branding IIT "pseudoscience" for outrunning its evidence, an unusually public rupture that says as much about the field's nerves as about the theory.

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) was a British mathematician and logician, the founder of theoretical computer science and one of the founders of artificial intelligence. In a single 1936 paper he gave the first precise mathematical account of computation. During the Second World War he did decisive codebreaking work at Bletchley Park against the German Enigma cipher. In 1950 he reframed the question of machine intelligence in terms that have governed the debate ever since. Prosecuted in 1952 for homosexuality, then illegal in Britain, and stripped of his security clearance, he died in 1954 of cyanide poisoning, generally judged a suicide. He received a royal pardon in 2013. His ideas sit at the headwaters of nearly every question this site raises about machines and mind.

Major Works

- "On Computable Numbers, with an Application to the Entscheidungsproblem" (*Proceedings of the London Mathematical Society*, 1936) — the Turing machine, computability, and the universal machine
- "Computing Machinery and Intelligence" (*Mind*, 1950) — the imitation game (the "Turing test")

Key Positions

The Turing machine and computability. Turing analyzed *computation* into its barest elements (a tape, a finite set of states, rules for reading and writing symbols) and used the resulting abstract machine to make "effectively calculable" mathematically precise. The analysis underwrites the **Church–Turing thesis**: the functions computable by an idealized mechanical procedure are exactly the Turing-computable ones. It is a claim about which functions can be mechanically computed — not, on its face, a claim about minds. See *Computation and the Church-Turing Thesis*.

The universal machine. Turing showed that a single machine, given a description of any other, can simulate it. The universal Turing machine is the theoretical ancestor of the stored-program

computer (one device that runs any program) and the formal root of the idea that the same computation can be realized on many different physical substrates.

The imitation game (the Turing test). Facing the question "Can machines think?", Turing judged it "too meaningless to deserve discussion" and proposed replacing it with an operational substitute. A human judge, communicating by text alone with a hidden human and a hidden machine, tries to tell which is which. A machine that reliably passes, one no judge can distinguish from the human by conversation, should, Turing argued, be credited with thinking. The move is deliberately behavioral: it sets aside what is going on inside in favor of what the system can do. Crucially, Turing **explicitly bracketed consciousness**. Replying to "the argument from consciousness," he declined to settle whether the machine has inner experience, treating that as a separate question from whether it thinks. He also anticipated and answered a battery of objections: the theological objection, the mathematical objection from Gödel, and "Lady Lovelace's objection" that a machine can only do what it is told.

Legacy. Turing's framing set the terms for everything that followed. Functionalism and the computational theory of mind inherit his universal-machine picture of substrate-independent organization. Searle's Chinese Room runs the imitation game in reverse: a case built to show that passing a behavioral language test, manipulating symbols by rule, is not sufficient for understanding. Contemporary work on large language models has revived the test directly. Systems now routinely produce human-indistinguishable text, which sharpens rather than settles the question Turing set aside.

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 (b. 1950), British-born, has spent most of his career at the University of Texas at Austin, where he is Distinguished University Chair, after earlier posts at Temple University and King's College London. He is the leading contemporary defender of strong representationalism, the view that phenomenal consciousness consists in a certain kind of representational content. Three decades of work built that position into its most sophisticated form. Then, in a 2021 book whose argument he restated in 2024, came a late and much-discussed conversion to panpsychism.

Major Works

- *Ten Problems of Consciousness* (1995) — the foundational statement of PANIC theory; the full transparency argument; Mary's Room handled representationally
- *Consciousness, Color, and Content* (2000) — PANIC theory refined; Eloise and the transparency argument in its sharpest form
- *Consciousness and Persons* (2003) — the ego theory vs. bundle theory; personal identity; the extension of representationalism
- "Representationalism and the Transparency of Experience" — the canonical transparency paper
- "Homunculi Heads and Silicon Chips" (2011) — the historical thesis; why phenomenal character depends on causal history
- "What Uninformed Mary Can Teach Us" (2004) — the knowledge argument answered via representational capacity
- *Tense Bees and Shell-Shocked Crabs* (2017) — animal consciousness and the scope of phenomenal representation
- *Vagueness and the Evolution of Consciousness: Through the Looking Glass* (2021) — the late turn: consciousness, vagueness, and the panpsychist conclusion

Key Positions

PANIC theory (the 1995 framework). In *Ten Problems of Consciousness*, Tye's influential early version of strong intentionalism identifies phenomenal character with **Poised, Abstract, Nonconceptual, Intentional Content**. *Poised*: the content stands ready for cognitive use, at the interface between sensation and downstream reasoning. *Abstract*: the content represents properties abstractly, not particular instances tied to specific objects. *Nonconceptual*: the subject need not possess concepts for the represented properties. *Intentional*: the content is world-directed, with accuracy conditions. On that framework, phenomenal character just *is* PANIC content — not something correlated with or merely realized by it. Tye later called PANIC his "old view" and denied that it was equivalent to his newer externalist account, so the book treats PANIC as a historically important hypothesis rather than his unrevised current position.

The transparency argument. Tye's deployment of transparency in "Representationalism and the Transparency of Experience" is the most rigorous available. The argument starts from a phenomenological observation: introspect your visual experience and you find the world's properties, not the experience's properties. From there it moves through the transparency thesis to the conclusion that phenomenal character consists in representational content. The argument engages Boghossian and Block's objections directly. See *Transparency*.

The historical thesis. "Homunculi Heads and Silicon Chips" (2011) argues that history matters crucially to phenomenology: two individuals intrinsically identical at a time can differ in phenomenal character because of what was going on in them in the past. The China-body system (billions of people replicating a brain's causal organization) lacks phenomenal states. The problem is not organizational complexity; the system lacks the developmental history through which that organization acquired world-tracking content. This bridges intentionalism and embodiment: phenomenal character depends on causal-historical relations through which content gets fixed.

Mary's Room via representational capacity. "What Uninformed Mary Can Teach Us" (2004) shows how to answer Jackson's knowledge argument without conceding non-physical facts. What Mary gains on release is a *new representational capacity* (the ability to represent redness through direct perceptual engagement), not new knowledge of a previously hidden inner property. The knowledge argument, on this diagnosis, trades on a confusion between knowing that p and being able to represent red directly.

The late conversion to panpsychism. In *Vagueness and the Evolution of Consciousness* (2021), restated in 2024, Tye announced a change of course. The driving argument runs through vagueness: consciousness cannot be a vague matter (nothing can be borderline conscious), yet every gradualist evolutionary story requires consciousness to have arrived by degrees. Tye's resolution makes experience, or a proto-form of it, a fundamental and pervasive feature of the natural world. This is a significant revision of his earlier framework, not a minor clarification.

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) was a Hungarian-American mathematician and polymath whose work reshaped logic, quantum mechanics, economics, and computing. He gave the stored-program computer its canonical design (the "von Neumann architecture" that nearly every machine still follows) and, at the end of his life, turned that engineering insight on the brain itself. His short, unfinished final book asked directly whether the computer and the brain are two species of one kind of thing, and gave an answer more careful than the question's later enthusiasts.

Major Works

- *Mathematical Foundations of Quantum Mechanics* (1932) — the rigorous formal framework for quantum theory
- "First Draft of a Report on the EDVAC" (1945) — the stored-program architecture: instructions and data held in one memory
- *Theory of Games and Economic Behavior* (1944, with Oskar Morgenstern) — the founding of game theory
- *The Computer and the Brain* (1958, posthumous) — a sustained comparison of digital machines and the nervous system

Key Positions

The stored-program computer. Von Neumann's design holds a machine's instructions in the same addressable memory as its data. A computer can therefore treat its own program as something to read, modify, and run. This is the practical realization of Turing's universal machine and the reason one device can run any software.

Self-reproducing automata. He proved that a machine could, in principle, contain a complete description of itself and use it to build a copy. In the abstract, this anticipated the logic later found in biological reproduction.

The brain as a partly-digital, partly-analog organ. In *The Computer and the Brain*, von Neumann pressed the computer analogy hard and then, tellingly, pulled back from it. The neuron's all-or-none firing looked digital. But the brain, he argued, runs on a "language" of low logical depth and high parallelism, tolerant of error and statistical in character, unlike any digital computer of his day. He concluded that whatever the brain's logic is, it is *not* the logic of the machines he had built.

Legacy. Von Neumann supplies both halves of a tension that runs through all later debate about machines and minds. His architecture is the engineering root of the idea that the same computation runs on many substrates, the picture that makes functionalism and machine-mind talk feel natural. Yet his own analysis of the brain resisted the easy identification, insisting that

the nervous system computes in a way fundamentally unlike a digital processor. The founder of the stored-program computer was, on the mind, a substrate skeptic.

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) left Vienna for Cambridge, held its Chair in Philosophy from 1939 to 1947, and produced two distinct and largely opposed philosophical systems: the early *Tractatus Logico-Philosophicus* (1921) and the later *Philosophical Investigations* (1953, posthumous). The later work, with its investigations into language, rule-following, private language, and the nature of mind, has influenced contemporary philosophy of mind more than almost any other single text.

Major Works

- *Tractatus Logico-Philosophicus* (1921) — the picture theory of meaning; logical atomism; the limits of language
- *Philosophical Investigations* (1953, posthumous) — the rule-following considerations; meaning as use; the private language argument; the dissolution of philosophical problems
- *The Blue and Brown Books* (1958) — preparatory lectures toward the *Investigations*
- *On Certainty* (1969, posthumous) — epistemic certainty and the foundations of doubt
- *Zettel* (1967, posthumous) — collected fragments

Key Positions

Meaning as use. The methodological principle of the *Philosophical Investigations* (sec. 43): "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." Words do not have meanings as inner attached entities or labels for objects; they have uses in public practices. This overturns the

picture theory of meaning from the *Tractatus* and relocates meaning from inner mental states to shared, public linguistic activity.

The rule-following considerations. The *Investigations* (secs. 138–242) investigates what it means to follow a rule: to use a word "the same way" on successive occasions. Wittgenstein argues that no finite set of past applications of a rule determines what counts as following it in a new case; any new application is compatible with indefinitely many rules. What makes one application "correct" is not anything in the individual's inner states but the practice of correction and agreement within a community. Following a rule is a social achievement, not a solitary inner event.

The private language argument. The *Investigations* (secs. 243–315, especially sec. 258) is the most discussed anti-Cartesian argument of the twentieth century. Wittgenstein imagines someone who tries to define a term ("S") purely by private, inner ostension: pointing, in imagination, to a private sensation. His argument: the attempt fails. Without a criterion of correctness (a standard against which one could determine whether the next use of "S" agrees with the first) there is no fact of the matter about whether the word is being used consistently. A "private language" would have no standard of correct use and therefore no meaning.

Dissolution rather than solution. Wittgenstein's therapeutic methodology: many philosophical problems arise from misunderstandings of how language works, from taking grammatical surfaces at face value, from picturing mental states as objects, from importing one domain's grammar into another. Philosophy's job is to dissolve these confusions by tracing the actual use of expressions, not to propose new theories.

The assault on the Cartesian inner standpoint. Wittgenstein's later work is a sustained critique of the picture that the mind is a private realm to which only the subject has direct access. The private language argument, the rule-following considerations, and the treatment of psychological concepts all converge on the conclusion that the Cartesian picture misunderstands the grammar of mental-state attributions.

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.

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

What Reification Is

To *reify* something is to treat it as a thing (a particular, individuated, persisting object) when it in fact names a pattern, a process, a relation, or a capacity. Reification turns descriptions into entities. It promotes a way of talking about phenomena into a claim that those phenomena consist in inner items the talk picks out.

Examples from outside philosophy of mind make the move easier to see. *Inflation* is a pattern in prices over time; treating it as a thing that "acts" on the economy is a useful shorthand that becomes misleading when pushed. *Traffic* is what cars are doing; "the traffic decided to" treats the pattern as an agent. *Mind* is what minded creatures do; treating it as an inner organ with properties of its own repeats the move where the stakes run higher.

In philosophy of mind, reification is the engine of most of the bad pictures an anti-reificationist program has to clear away. Terms like *mind*, *meaning*, *intelligence*, *information*, *consciousness*, *qualia*, *intentionality* function as shorthand for capacities, processes, or relations. When they get reified, with *mind* promoted to an inner substance, *qualia* to inner objects, and *consciousness* to a stuff or a property of an inner item, philosophical problems multiply that wouldn't exist without the reification.

The Diagnostic Methodology

Anti-reification is not a thesis but a methodology. It governs how an anti-reificationist talks about mental phenomena, what kinds of explanations it accepts, and what kinds of philosophical problems it treats as artifacts of bad pictures rather than as discoveries about nature.

The methodology has three working rules.

1. Cash out what process or relation a term abbreviates. Confronted with *intentionality*, the question is not "what kind of thing is it" but "what relation does the term name." Confronted with *consciousness*, ask what a conscious system actually does, undergoes, or relates to that systems lacking consciousness don't. The point is not eliminativism (the phenomena are real) but specifying what they consist in without turning them into entities.

2. Name the level-confusion when diagnosing a reification. Reifications have structures; they don't happen at random. The most common structures the methodology diagnoses:

- **Simulation vs. realization.** Treating a description of a process as the process itself. See *Simulation and Realization.*
- **Syntax vs. semantics.** Treating formal symbol manipulation as meaning. See *The Chinese Room.*
- **Description vs. thing described.** Treating ways of talking about phenomena as discoveries about their inner structure.

- **Vehicle vs. content.** Treating the physical thing doing the representing as the thing represented. See *Vehicle and Content*.
- **Observer-relative vs. intrinsic.** Treating properties that systems have only relative to interpreters as intrinsic features.

Each is a different way of pulling a description into the world it describes. Naming which level-confusion is operating lets the diagnosis land.

3. Articulate identities rather than leaving them flat. "Pain is C-fiber firing" overcommits to one neural type and hides the explanatory structure. "Within a conscious state borne by an identified subject, pain's phenomenal character consists in a damage-presenting representational profile under a bodily-affective mode, realized neurally" states the book's identity claim while specifying its relata and realization. Here *consists in* expresses identity in reader-facing prose, not a weaker relation of dependence or constitution.

This is the discipline at work: use verbs with care. *Consists in* states the identity claim; *depends on* marks dependence; *represents* names content; *realizes* names implementation; *functions as* names role; *arises from* names causal history. They are not interchangeable.

What Anti-Reification Is Not

The methodology is not eliminativism. It does not hold that consciousness, qualia, intentionality, meaning, or mind don't exist. It holds that these terms name patterns, processes, relations, and capacities rather than inner substances, and that philosophical problems generated by treating them as inner substances dissolve when the right level of description is restored.

Nor is it deflationism in the technical sense, the claim that all mental talk is "merely useful talk" with no underlying reality. The underlying reality remains fully real: the brain doing what brains do, embedded in environments doing what they do, with the relational and representational structure that constitutes mental life. Only the inner-substance reading of the vocabulary goes.

Nor does the methodology refuse abstractions. An anti-reificationist still talks about *the mind*, *consciousness*, *the inner voice*, *qualia* — including in passages that criticize those very terms — because the abstractions are useful shorthand for the patterns they name. The methodology requires only that the shorthand not be promoted to a metaphysical claim about inner items.

Applying the Methodology

Anti-reification pairs naturally with physicalist monism, strong representationalism, and embodied realization, and it shapes how each gets articulated.

On consciousness. Treat consciousness as what a system *does* when it represents the world in certain ways, not as a stuff or substance the system contains. Phenomenal character is how the representing is, not what the representing has inside it. See *Intentionalism*.

On qualia. Reject the picture in which qualia are intrinsic, non-relational, inner properties of experiences. The redness of red-experience is a way of representing redness as a property of surfaces, not an inner item with the property of being red. See *Qualia*.

On the inner voice. The inner voice is not a private inner thing but public speech run silently — the same activity, with the same meanings, drawn from the same shared language. Reifying it as a private chamber produces the inner-theater problems an anti-reificationist program exists to dissolve. See *Inner Speech*.

On information and computation. *Information* names a measurable relation between physical states; it is not a stuff that flows. Unconstrained computational mappings remain observer-relative, but disciplined, counterfactual-supporting causal-computational organization can constitute an intrinsic physical fact. Intrinsic implementation alone still fixes neither semantic content nor mentality.

On the mind itself. *The mind* names what minded creatures do — perceiving, representing, remembering, intending, reasoning, feeling. Treating the mind as an inner thing distinct from those activities creates a recurrent source of problems in philosophy of mind. Many of this book's central targets inherit that mistake, though not every problem reduces to it.

The Phenomenology Caveat

A standard objection: anti-reification denies the obvious phenomenology. When I have a sharp pain, the pain feels like a thing. When I imagine a red apple, the image seems to be there in some sense. Doesn't anti-reification have to deny these reports?

It does not. The phenomenological reports are real and accurate, and so is what they report — there really is a sharp pain, there really is an imagined apple. The anti-reification move is not to deny that these phenomena occur, but to deny that they occur as inner items presented to an inner observer. The pain is a representational state of the body, the apple-imagery is a representational activity directed at a (non-existent or possible) apple. The phenomenology survives; only its inner-object interpretation goes. That answer keeps the methodology from sliding into the eliminativism it sometimes gets confused with.

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 is the view that mind consists in behavior: to be in pain, to believe it will rain, to want coffee just is to behave, and to be disposed to behave, in certain characteristic ways. For roughly the first half of the twentieth century it dominated psychology and much of philosophy, and understanding why it rose and why it fell explains most of what came after it in the philosophy of mind.

The Problem It Was Solving

Descartes left philosophy with a big question: if our mind and body are separate, how do they work together? And how could mental states, inaccessible to third-person observation and unverifiable in principle, ever be studied scientifically? Dualism made psychology look like ghost-hunting.

Behaviorism's response was radical: eliminate the inner entirely. Mental states aren't private inner events — they're patterns of observable behavior and behavioral dispositions. To say someone is in pain isn't to report an internal event; it's to say they are disposed to wince, withdraw, say "ouch," and seek relief. Pain just *is* that cluster of dispositions. No ghosts, no inner theater, no verification problem.

This was a serious philosophical program, not a naive one. Its main forms:

- **Methodological behaviorism** (Watson, Skinner): psychology should restrict itself to observable behavior as its subject matter, whatever the metaphysical facts about inner states
- **Logical behaviorism** (Carnap, Hempel): mental vocabulary can be *translated* into behavioral vocabulary without remainder — "pain" means nothing more than the relevant behavioral dispositions
- **Philosophical behaviorism** (Ryle): Cartesian talk of inner mental events rests on a category mistake — what looks like a report of an inner state is actually a description of behavioral competence and disposition

Ryle's *The Concept of Mind* (1949) is the philosophical high-water mark. His "ghost in the machine" phrase remains the most famous diagnosis of the Cartesian picture: dualism, Ryle argued, treats the mind as a ghostly engine hidden inside the physical body, causing behavior from the inside.

The Standard Objections

The objection that did the most damage: behaviorism seems unable to distinguish genuine mental states from perfect behavioral mimicry.

Consider pain. A person in genuine pain and a skilled actor pretending have matching behavioral dispositions in the staged circumstances: both wince, withdraw, say "ouch," seek relief. Yet one hurts and the other doesn't. Hilary Putnam pressed the point in both directions, imagining "super-Spartans" who feel pain but have trained away every pain-behavior: if behavior can occur without the feeling and the feeling without the behavior, neither can define the other.

Behaviorists had a reply. Dispositions reach wider than any performance: offstage, in unstaged circumstances, the actor's dispositions diverge from the sufferer's, and even a super-Spartan stays disposed to avoid what damages him. The critics' rejoinder: stipulate the pretense or the stoicism total, and a felt difference still seems to remain, which suggests the feeling never consisted in dispositions at all.

A second objection targets **the circularity of dispositional analysis**. To cash out the disposition to "seek relief," you need to invoke the belief that relief is available, the desire to obtain it, the intention to act. But beliefs, desires, and intentions are themselves mental states awaiting behavioral analysis, and the circle never closes. Roderick Chisholm and Peter Geach both pressed

this point in the 1950s: no dispositional analysis of one mental state completes without invoking others.

What Survived It

Even behaviorism's harshest critics kept two of its lessons. First, Ryle's diagnosis outlived the analysis it served: much philosophical confusion about mind does come from treating mental-state vocabulary as if it named hidden inner objects, and "the ghost in the machine" remains the standard name for that picture. Second, mental states need some anchor in the publicly accessible world. Behaviorism chose behavior; later theories chose causal roles or world-involving relations. Different answers to the same demand.

The Transition to Identity Theory

By the late 1950s, the objections had persuaded most philosophers that inner states could not be analyzed away. U. T. Place (1956) and J. J. C. Smart (1959) drew the conclusion: if we need inner states, and dualism is unacceptable, then inner states must be physical — specifically, brain states. Pain is C-fiber firing, as the slogan went. This is identity theory: the inner is real, and it's neural.

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

Carbon chauvinism is the view (most often raised as a charge against a position rather than a banner anyone marches under) that mind or consciousness can be realized only in carbon-based biological matter, so that no system built from other materials, however organized, could ever genuinely think or feel. The phrase carries over from astrobiology, where Carl Sagan used "carbon chauvinism" for the parochial assumption that life elsewhere must be carbon-based because life here is. Transplanted into philosophy of mind, it names the parallel assumption about minds: that the *stuff* a system is made of, and not merely the way that stuff is organized, is what decides whether anyone is home.

The label marks one side of a genuine question — the **substrate question**: does having a mind depend on being made of a particular kind of material, or only on standing in the right organization and relations, whatever the material? To call a view "carbon chauvinist" is to accuse it of answering "a particular material" without adequate reason — of mistaking a feature of the only minds we happen to know for a requirement on minds as such.

The Charge and Its Place in the Debate

The term belongs to a family of objections in the functionalism literature. Ned Block, in his diagnosis of functional theories of mind, draws the standard pair: a theory of mind errs toward **chauvinism** when it withholds mentality from systems that have it (by tying mind too tightly

to the specific details of human or biological make-up), and toward **liberalism** when it grants mentality to systems that lack it (by tying mind to organization so abstract that economies and populations would qualify). Carbon chauvinism is the substrate-keyed species of Block's chauvinism: the error of building one's own material into the definition of mind.

The standard pressure against carbon chauvinism comes from **multiple realizability** — the observation, central to functionalism, that the same mental state can be realized in different physical substrates (Putnam, Fodor). If pain is a functional state, a role defined by its causes and effects, then what matters is whether a system occupies the role, not what occupies it. Octopus, Martian, and (in principle) machine could all be in pain, because being in pain is a matter of organization, not chemistry. On this view the demand for carbon looks like a category mistake: it confuses the realizer with the role.

The Case for Substrate Dependence

The opposing side need not rest on mere prejudice, and its more careful versions resist the "chauvinist" label.

Searle's biological naturalism holds that consciousness is caused by specific neurobiological processes and realized in brain structures, the way digestion is caused by specific biochemical processes in the gut. On this view consciousness has real causal-biological preconditions, and a system lacking the relevant causal powers (for instance, one that merely *simulates* the formal organization of a brain) would not thereby be conscious. Searle is explicit that this is *not* carbon chauvinism: his claim is not that only carbon could ever do it, but that whatever does it must have the right causal powers, and that we know brains have them while formal organization alone does not guarantee them. The real target of his Chinese Room is not silicon but *syntax* — the thesis that symbol manipulation is sufficient for understanding (see *The Chinese Room*).

Type-identity theory is substrate-committed in a different way: if a mental state type just *is* a physical state type (pain is C-fiber firing), then a system without that physical type lacks that mental state. Multiple realizability is the standard objection to exactly this commitment.

A third strand turns on the **simulation/realization distinction**: a perfect computational model of a process is not always an instance of it — a simulated hurricane leaves the ground dry. If consciousness is among the processes that must be *realized* rather than merely *modeled*, then running the right program in any substrate would not be enough; the substrate would have to do the relevant causal work, not just represent it (see *Simulation and Realization*).

The Argument from Organizational Invariance

The strongest argument *against* carbon chauvinism is Chalmers's pair of thought experiments, the **fading qualia** and **dancing qualia** scenarios, supporting a principle of organizational invariance: any two systems with the same fine-grained functional organization have qualitatively identical experiences. Imagine replacing your neurons one at a time with silicon chips that preserve every functional input-output relation. If experience depended on the carbon, it would have to either fade gradually or wink out at some threshold — yet by hypothesis nothing functional changes, so you would notice nothing and report nothing, which Chalmers argues makes "fading" or "dancing" qualia incoherent. The conclusion: experience tracks organization, not the material

that implements it. Tye presses a similar point with his homunculi-heads and silicon-chip cases. Chalmers is no physicalist, yet he uses the argument to defend substrate independence; the issue cuts across the physicalist/dualist divide.

Two Questions Often Run Together

A recurring source of confusion: rejecting carbon chauvinism (the substrate need not be carbon) is a different move from accepting that *organization or computation alone* suffices for mind. One can hold that the material is not what matters while still denying that mere formal organization, a program run on any hardware, is enough. The ground for that denial: what matters is the right *world-involving causal* organization (a body, an environment, a causal history), realizable in more than one material but not secured by formal symbol manipulation alone. So "is it made of the right stuff?" and "does running the right program make a mind?" are distinct questions; a view can answer *no substrate requirement* to the first and still answer *no* to the second. This is the space most contemporary accounts of artificial minds try to occupy — between carbon chauvinism on one side and the liberalism that grants a mind to anything that passes a behavioral test on the other.

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 the claim that running the right computer program suffices for genuine understanding. Published in "Minds, Brains, and Programs" (1980), it remains the most famous argument in the philosophy of artificial intelligence, and nearly every position in the field has defined itself partly by how it answers the Room.

The Argument

The setup:

Imagine a person locked in a room. Slips of paper with Chinese symbols are passed in. The person consults an enormous rulebook that specifies, for any input sequence of symbols, exactly which output symbols to produce. The person follows the rules, produces the outputs, and passes

them back. From outside, the room produces perfectly appropriate Chinese responses to Chinese questions. It passes the Turing Test for Chinese.

But the person inside understands no Chinese whatsoever. They manipulate symbols by shape rather than meaning. The setup therefore establishes operator-level formal rule-following, not understanding by the implemented whole. The Chinese tokens retain the public, derived meaning supplied by the linguistic practice that produced them.

Searle's conclusion: Syntax is not sufficient for, and not constitutive of, semantics. No amount of formal symbol manipulation, no matter how sophisticated, produces genuine understanding, intentionality, or meaning. Therefore strong AI, the thesis that running the right program is sufficient for having a mind, is false.

The Systems Reply

The most common objection, which Searle credited to critics at Berkeley: the *person* doesn't understand Chinese, but the *system* does — the person plus rulebook plus room as a whole. The man, on this reply, plays the role of a single component, a CPU or one neuron among billions, and nobody ever claimed that a component understands. Understanding belongs to the whole; the intuition that it must live in the man is precisely the mistake.

Searle's reply: internalize the system. Memorize the rulebook, learn to carry out the operations in your head while walking around outside. Now the system has no parts beyond you — the system *is* you — and you still understand no Chinese. On Searle's diagnosis, the systems reply relocates the mystery without explaining it.

The Robot Reply

Put the program in a robot instead. Give it cameras and motor control; let it move through the world its symbols describe, name what it sees, act on what it hears. Meaning, on this reply, never comes from symbol-to-symbol relations alone — it comes from symbol-to-world connections, and the robot has them.

Searle's reply comes in two parts. First, the robot reply concedes the thesis under attack: if world-involving causation is needed, then running the right program was never sufficient by itself. Second, it doesn't help. Put the man in the robot's control room and the camera feeds arrive as still more uninterpreted symbols; he goes on matching shapes. Transducers bolted to the outside add no understanding on the inside.

The Brain Simulator Reply

Suppose the program simulates the actual neuron-by-neuron behavior of a Chinese speaker's brain. Surely then it understands?

Searle's reply: simulation is not realization. A perfect simulation of a storm does not make you wet. A perfect simulation of digestion does not digest anything. A perfect simulation of a brain's causal processes does not thereby instantiate those causal processes — it instantiates a *description* of them.

"Is the Brain a Digital Computer?" (1990)

Searle's companion paper sharpens the critique. Computation is not an intrinsic feature of physical systems — it is observer-relative. We *assign* computational descriptions to physical processes. The brain is not, intrinsically, a digital computer; it is a biological causal system, and attributing computation to it requires an interpreter. But genuine intentionality (meaning, understanding, reference) is intrinsic, not observer-relative. Therefore, Searle argues, computational descriptions cannot explain intentionality.

This argument supplements the Chinese Room: on Searle's account, syntax not only fails to generate semantics, but computation itself fails to be the right kind of physical fact to generate anything intrinsic. The point blocks the response that "we just need more complex computation."

Application to Large Language Models

The Room now frames the central dispute about modern AI, but the analogy needs a system boundary. A bare language model predicts tokens from learned patterns in training data. A deployed product may also include persistent memory, retrieval, tools, images, live feeds, and action loops. Training can give internal mechanisms causally active derived content, and corpus-mediated history may transmit public reference. These achievements exceed the hand-written Room without yet establishing understanding by the containing system.

Critics correctly insist that operator ignorance proves nothing about the whole. The live question asks whether content-bearing processes form one persisting, world-answerable economy of memory, correction, inference, planning, learning, and action. A sufficiently integrated computational implementation could pass that test. The Room defeats the shortcut from formal success to understanding; it does not settle every system capable of running a program.

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 is one of the foundational results in the theory of computability. In its standard formulation, it identifies the intuitive notion of *effective calculability* (what can be computed by a systematic procedure) with what a Turing machine can compute. Proposed

independently by Alonzo Church and Alan Turing in 1936, it has shaped computer science, mathematics, and philosophy of mind ever since.

The thesis gets invoked constantly in arguments about the nature of mind. If the brain is a physical system, and physical processes are in principle computable, then, the inference goes, the brain is a Turing machine, and any sufficiently complex computer running the right program would have a mind. That inference moves through several steps, and each deserves separate examination: the Church-Turing thesis itself, the further claim that minds are computational systems (computationalism), and the claim that computational processes can generate semantic content (the syntax-semantics question).

Keeping these three apart matters for evaluating any argument about artificial intelligence, machine understanding, or the computational theory of mind.

Three Distinct Claims

1. The Church-Turing thesis (CTT)
The CTT, in its standard form, identifies the intuitive notion of *effective calculability* with what a Turing machine can compute: any function computable by an effective procedure is Turing-computable. It is a claim about the *extension of "computable,"* not a claim about minds, brains, or meaning. A separate **physical** Church-Turing thesis (Cotogno, Galton) asks whether every physically realizable process is Turing-computable — a stronger and more contestable claim, bearing on hypercomputation and the limits of physical realizability. Even granted in full, physical version included, the thesis says nothing about understanding on its own. It tells you a function can be *computed*; it stays silent on whether computing it amounts to *thinking* it.

2. Computationalism
Computationalism is the further thesis that cognition consists in computation — that the mind operates as, or gets implemented by, a computational system. **Computational functionalism** places that thesis inside the broader functionalist framework: functionalism identifies mental states by their causal roles, while computational functionalism adds that the relevant roles are computational states. Strong computationalism goes further again and claims that running the right program suffices for mentality. None of those additions follows from functionalism alone. The computational picture earned its long run at the center of cognitive science honestly. Turing had shown how a purely mechanical process can carry out any effective procedure, which made it possible, for the first time, to see how a physical mechanism might reason without a little person inside doing the reasoning. Hilary Putnam built machine functionalism on the idea; Jerry Fodor built a theory of thought on it; Ned Block's "The Mind as the Software of the Brain" gives the picture its clearest statement.

The crucial logical point, stressed by Gualtiero Piccinini among others: computationalism does **not** follow from the CTT. "The brain's processes are computable" and "the brain is a computer (and the mind its software)" are different claims; the first can be true while the second is false. Critics argue that running the two together lends computationalism a borrowed air of mathematical security — the security of a thesis that was never about minds.

3. Syntax and semantics

The third claim is Searle's: a computational process is defined over *syntax*, formal symbol manipulation, and syntax does not, by itself, yield *semantics*. Running the right syntax is not thereby meaning anything. Searle sharpens this with the observation (Negru) that *syntax itself is observer-relative*: being a "symbol" or a "computation" is not an intrinsic physical feature of a system but something assigned to it by an interpreter — which is exactly what a mind with original intentionality would have to supply, and so cannot be what computation provides. See *John Searle* and *Intentionality*.

A Related Distinction: Simulation and Realization

One further distinction runs through this territory. A perfect computational *simulation* of a process is not always an *instance* of that process: a simulated hurricane makes nothing wet, and a simulated digestion digests nothing. Whether a faithful simulation of a brain would realize what the brain realizes, or merely model it, is a separate question from anything the CTT settles. See *Simulation and Realization*.

The Live Debate

The most direct recent challenge to Searle's side comes from Vladimír Havlík's "Meaning and Understanding in Large Language Models" (2024), which argues that the assumption of a *complete* gap between syntax and semantics is unjustified — that the strong claim ("no syntactic process can produce semantics") does not follow from the Chinese Room alone, and that "syntactic semantics" (Rapaport) may bridge the gap. The strong/weak distinction marks the live edge of the debate: whether the gap is total, and what argument could show it.

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

Some things mean what they do because a user or practice lends them meaning. Other states have accuracy conditions that do not depend on anyone interpreting them that way. John Searle calls the first *derived* intentionality and the second *original* or *intrinsic* intentionality. The distinction concerns dependence on interpretation, not possession by an owner.

A stop sign really means *stop*, but it means that through a public practice. Remove the practice and the metal retains its shape and causal powers but not its instruction. By contrast, if a frog's producer-consumer circuitry was stabilized because one state guided prey-directed behavior under one condition and another state under another, that mapping may help explain what the states can get right or wrong without waiting for an observer to assign them meanings.

Three Questions

Artificial systems make three questions especially easy to confuse:

1. **Content fixing:** Does the proposed mapping explain the mechanism's stabilization and survive interventions on its producers, consumers, and success conditions, or is it merely a convenient gloss?
2. **Level of attribution:** Is the content properly attributed to a component, a coupled mechanism, or the whole system?
3. **Scope of capacity:** Does content in one mechanism support perception, belief, understanding, planning, or consciousness at the level of the larger system?

An affirmative answer to the first question does not settle the other two. A subpersonal state can carry content without becoming a tiny subject. A system can contain content-bearing mechanisms without thereby understanding their content.

To consider a robot as possibly having original content with a subject, both the content-fixing and attribution questions need independent support. The relevant history and producer-consumer use must fix a world-involving correctness standard not exhausted by an interpreting practice. An inherited task does not by itself decide whether they do. Interventions must also locate one relatively maximal, persisting control organization in which those states update a shared economy of perception, expectation, memory, correction, planning, and action. The first finding makes original content a live attribution; the second identifies its candidate subject. Neither finding establishes consciousness, and the second does not create the first.

Original Does Not Mean Uncaused

Original content need not arise from nowhere. Natural selection, individual learning, and perhaps artificial training can establish producer-consumer mappings whose semantic interpretation does explanatory work. *Original* concerns whether history and use fix accuracy conditions not exhausted by an interpreting practice. It does not require freedom from causal or semantic inher-

itance. A training process can make an inherited interpretation causally indispensable without making it independent; inheritance alone cannot show that independence is absent either.

The positive test matters. A representational description earns realism when the mapping helps explain why the mechanism has the organization it does, predicts its success and failure, and remains privileged across relevant interventions. If rival mappings perform equally well throughout, the content may remain indeterminate at that grain. If the mapping merely redescribes what an engineer intended the device to do, the content remains derived.

Original Does Not Mean Novel

A second confusion travels under the same word, and it costs more arguments than the first. In ordinary usage, *original content* names something no one has produced before — a new sentence, a new image, a new tune. Language models produce that in quantity, and some of it repays attention. Nothing in this distinction denies it.

The philosopher's sense asks a different question. Not whether the output has appeared before, but what fixes the standard that makes it right or wrong. Novelty belongs to the product; originality in this sense concerns the standard's dependence on interpretation. A system can produce unprecedented output indefinitely while every standard it answers to remains dependent on an interpreting practice, and a system can produce dull, repetitive output whose accuracy conditions its own history and use fix. The two properties vary independently.

Keeping them apart matters because the popular sense makes the philosophical claim sound like a denial of the obvious. It denies nothing about what these systems can write.

The LLM Case

An emitted word carries public, derived meaning as an English token. Inside a language model, training produces elaborate geometry, learned sensitivities to textual distributions, and states that make indispensable causal differences to later computation. Fintan Mallory applies consumer-based teleosemantics to word2vec and argues that training can give its embeddings original content. His claim concerns actual word-types and the contexts in which they occur, not the numerical vectors themselves or the objects the words name.

On Mallory's account, an embedding combines a description with a directive: a word-type occurs in a particular context, and the downstream mechanism should produce a distribution over context words. Training links the two. Its updates respond to differences between predicted distributions and those in the data. The proposed content helps explain how producers and consumers acquired their organization; it does not rest on the geometry alone.

The external dictionary raises a question about which relations fix content, not a quick refutation. A coordinated change to word indices and the corresponding matrices can preserve the relation between actual words and their contexts. Changing the external pairing instead changes a candidate content-fixing relation. Neither operation alone establishes dependence on an interpreting practice. Human children inherit public standards too; both human and artificial learners require assessment through the history and use that make their states answerable to what they concern.

Linguistic distributions belong to the world too. A model can track facts about the word *maple* without tracking the maple outside the window. Changing the tree while holding the corpus fixed therefore tests a proposed reference to that tree, not the proposed content about word use. The distinction limits the claim without dismissing it. Where content does remain dependent on an interpreting practice, *derived* still does not mean inert, superficial, or optional: such content can operate inside a network and help explain what it does.

Nor does mechanism content establish whole-model understanding. That requires a further account of integration across memory, learning, inference, conflict resolution, planning, and action. Appropriate perceptual organization and poising would be needed for the book's strong representationalist account of consciousness. These are questions about what the larger system can do with content, not extra ingredients that turn content on.

Self-Maintenance and Autonomy

Reciprocal self-maintenance may help explain autonomy, agency, persistence, welfare, and the boundaries of an individual. It does not create accuracy conditions. Adding a self-repair loop to a content-bearing mechanism need not change what its states represent; removing the loop need not make their content derived. Ownership can still matter to whose belief, plan, or experience a state helps constitute. It does not mint the state's content.

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 in veridical perception, we stand in immediate contact with mind-independent physical objects and their properties. We see the apple, not a representation of the apple, not a sense-datum caused by the apple, not a neural image of the apple. The apple itself is the object of awareness.

The view sits in tension with a powerful philosophical tradition. The argument from illusion, the argument from hallucination, perceptual variation, the time-lag of light, and the neuroscience of perception have each been taken to show that what we really perceive directly is a mental

intermediary (a sense-datum or phenomenal image), with the external world known only by inference from these inner items. Any defense of direct realism has to answer all five.

What Direct Realism Requires

The view requires two commitments:

1. **Mind-independence**: the objects of perception exist and have their properties independently of being perceived. The apple is red whether or not anyone looks at it.
2. **Immediacy**: perceptual awareness of these objects is not mediated by awareness of a mental intermediary. We do not first encounter an internal proxy and then infer the external object.

What direct realism does *not* require: - Infallibility. A direct realist can cheerfully admit that perception sometimes misleads us. A stick looks bent in water; the stick is straight. What the direct realist cannot grant is that these cases show perception is *always* mediated by inner objects. The classical objections all try to force that concession; the intentionalist replies below resist it. - Simultaneous contact. Seeing a star eight light-years away is still direct perception of that star — of how it was when the light left. Temporal mediation is not the same as representational mediation via inner objects.

The Five Classical Objections and the Intentionalist Replies

1. The Argument from Illusion

A stick half-submerged in water looks bent. The Müller-Lyer lines appear unequal. The perceiver seems aware of a property (bentness, inequality) that the object doesn't have. If we perceive the object and its properties, what are we aware of when the object lacks the perceived property? The traditional answer: a sense-datum that genuinely possesses it.

Intentionalist reply: the experience *misrepresents* the stick as bent. No mental object is bent; the experience simply gets the stick wrong, the way a false sentence asserts something without conjuring a domain where that thing is true. The object of awareness remains the real, physical, straight stick, mischaracterized by the experience. Illusion is misrepresentation, not displacement of awareness from world to mind.

2. The Argument from Hallucination

A hallucinating subject experiences a pink elephant. No elephant exists. The experience can be subjectively indistinguishable from a veridical perception. If veridical perception is direct contact with an existing object, hallucination cannot be — there is no object. But the two experiences seem phenomenologically identical. Conclusion: in both cases, what we are directly aware of is a mental entity.

Intentionalist reply: hallucination is a representational state whose content is unsatisfied. Its phenomenal character consists in the complete intentional structure — content under a perceptual mode, with its determinacy, attention, and field organization — not in the presence of an external object or a coat of intrinsic mental paint. Hallucination and veridical perception can therefore share a way of presenting while differing in whether the content is accurate and appropriately caused.

3. The Argument from Perceptual Variation

The same round coin appears elliptical from an angle; the same white wall appears pink under red lighting. If we are directly aware of the object, and the object doesn't change, why does awareness vary?

Intentionalist reply: representational content is inherently perspectival. The experience represents the coin as *round-viewed-from-this-angle* — a specific relational content that changes as the viewing angle changes. This variation accurately tracks a changing relation between perceiver and object; it does not indicate the presence of varying mental entities. We always see from somewhere; perspectival content encodes this without requiring inner objects.

4. The Time-Lag Argument

Light from a star eight light-years distant takes eight years to reach our eyes. The star may no longer exist. If direct perception requires contact with objects as they currently are, no perception is ever truly direct.

Intentionalist reply: the content of the perceptual experience is about the star as it was when the light left — a temporally distal state of a mind-independent physical object. Perception of temporally past states is still direct perception in the sense that matters: no mental intermediary is involved. Memory, testimony, and inference are also about the past; the time lag no more makes perception indirect than a letter makes its reader indirect about the events described. What matters is the absence of a mediating inner object, not temporal simultaneity.

5. The Argument from Science

Neuroscience describes perception as the end product of a causal chain: photon → retina → optic nerve → lateral geniculate nucleus → visual cortex → conscious experience. At each link, what is causally proximate is a neural state. If the immediate cause of experience is a brain state, perhaps the immediate object of awareness is a brain state.

Intentionalist reply: causal proximity is not the same as perceptual awareness. The neural vehicle of experience is a brain state; the content of experience is about the external world. We are no more aware of our neurons than we are aware of the ink when reading, or the window glass when looking into a garden. Dretske: to say we are aware of brain states because they are the proximate causes of experience is like saying we read ink because ink is the proximate cause of reading. The causal chain *enables* perception without replacing its objects. The vehicle is neural; the content is the world.

The Vehicle/Content Distinction

The key to understanding intentionalist direct realism is the distinction between the *vehicle* of a representation and its *content*.

A photograph of the Eiffel Tower is a piece of glossy paper (vehicle) that depicts the Eiffel Tower (content). When you look at the photograph, you attend to what it depicts (the Tower, the Seine, the Parisian sky), not to the chemical composition of the paper. The vehicle is transparent to its content.

Perception's neural activity realizes the vehicle; phenomenology discloses the world as presented and can also disclose the intentional mode of presentation. Introspecting visual experience of the apple may disclose its perceptual mode, determinacy, attention, and field organization. It does not disclose the neural vehicle or an inner qualitative intermediary as an object of awareness.

This is the Transparency Thesis applied directly to the directness question. Introspection turns up the world and never a content standing in front of it — and content could not stand in front of anything. What an experience says about the world takes up no position between a perceiver and what it concerns. So content interposes no screen, not even an unnoticed one. Representing the world is what perceiving it consists in, never a step along the way.

The Screen-Off Objection

The most serious challenge to intentionalist direct realism comes from Bill Brewer and others: if representational content can exist without its object (as in hallucination), then content looks like an autonomous mental item, a thing standing between us and the world. To perceive via content, the objection runs, is to perceive indirectly.

Intentionalist reply: the objection treats content as an item and then asks where the item sits. Grant that first move and the second question turns unanswerable — an item between perceiver and world screens the world off, and an item that screens nothing off looks idle. So refuse the first move. What an experience says about the world never amounts to a thing occupying a position; the meaning of a sentence stands nowhere between a reader and what the sentence concerns. Content therefore sits neither in front of the apple nor beside it as a medium: the experience presents the apple by saying how things stand with it, and no second act of awareness intervenes. The phenomenology bears this out — perception runs outward, presenting the world and never itself — which is exactly what the transparency thesis claims.

Brewer's own alternative holds that external objects must be *constituents* of experience, not merely what it refers to. That secures directness in the veridical case, and its critics press the cost: hallucination becomes either impossible or a categorically different kind of state, with nothing phenomenal in common with seeing.

Varieties of Direct Realism

Direct realism comes in distinct varieties that take different positions on hallucination and illusion:

View	Direct object of awareness	Hallucination	Illusion
Naïve realism	The external object itself (as constituent)	Needs a separate account — no object is present	Via disjunctivist split
Disjunctivism	External object (veridical) / no common object (hallucination)	By dividing the cases into two kinds of state	Via disjunctivist split
Intentionalist direct realism	External world *as represented*	Inaccurate content	Misrepresentation
Indirect realism	Sense-data / phenomenal intermediaries	Inner object present either way	Inner object bears the property — at the price of directness

Intentionalist direct realism holds that experience has representational content; that content is transparent; and that in veridical perception we are directly aware of the world as represented. Directness, on this view, is preserved not by objects being constituents of experience but by nothing in the experience standing between perceiver and world.

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 in the philosophy of perception. Metaphysical disjunctivism denies that genuine perception and a subjectively matching hallucination share one fundamental experiential nature, though they may share true descriptions, neural causes, or downstream effects. These views occupy direct realist territory: they reject the sense-datum veil and hold that ordinary perception presents the world itself. Metaphysical disjunctivists include M. G. F. Martin (University of California, Berkeley), Paul Snowdon (University College London), and J. M. Hinton, who introduced the view in the 1960s. Epistemological disjunctivism, associated especially with John McDowell (University of Pittsburgh), advances a separate thesis about perceptual reasons and knowledge.

What Disjunctivism Claims

In Martin's metaphysical formulation, perception and hallucination share no fundamental experiential nature. They may still share true descriptions, neural causes, or downstream effects. The name comes from the resulting analysis of how things *seem*: "it seems to S that p" is true in virtue of a *disjunction* — either S genuinely perceives that p, **or** S is in a state merely indistinguishable from perceiving that p. The two disjuncts are not species of one fundamental experiential genus.

This is the explicit denial of what disjunctivists call the **highest common factor** picture: the idea that the most we are entitled to, given that hallucination is possible, is the inner state common to the good case and the bad case — and that this common state is what perception *really* delivers. The sense-datum theorist and the standard representationalist both accept some version of the common factor; they differ only on what it is (a sense-datum, a representational content). The disjunctivist rejects the framing wholesale: the good case is metaphysically *better* than the bad case, not the bad case plus an extra epistemic bonus.

Two metaphysical strategies recur in the literature, alongside a distinct epistemological thesis:

- **Positive metaphysical disjunctivism** (M. G. F. Martin): the veridical case is constituted by an *object-involving* relation — the perceived object is literally a constituent of the experience. Hallucination is a distinct state.
- **Negative metaphysical disjunctivism**: hallucination receives its characterization largely through its indiscriminability from perception rather than through a shared fundamental experiential nature.
- **Epistemological disjunctivism** (especially John McDowell): perceptual reasons in the good case can provide a kind of warrant unavailable in a subjectively matching hallucination.

The Pressure Point: Hallucination

The recurring objection concerns the positive nature and rational role of hallucination. Martin's negative answer does more than shrug: a hallucination indiscriminable from perception inherits inclinations to judge and act from the good case. Crane's pressure remains. Can that dependent epistemic relation explain how perception and hallucination serve as the same sort of subjective reason without a positive experiential feature common to both? Positive and moderate relational views distribute the costs differently, so no criticism of Martin's austere bad case refutes the family as a whole.

A rival reply, common-factor representationalism, holds that both perception and hallucination share representational content, and that direct realism is recovered at the level of content and transparency rather than constitution. On this account, the disjunctivist buys directness at the cost of an awkward theory of hallucination.

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 holds that mind and the physical are fundamentally distinct — that a complete account of the physical world leaves something out. It is one of the oldest and most persistently attractive positions in philosophy of mind, drawing strength from the apparent irreducibility of conscious experience to anything merely physical. In its various forms, it has been defended by Descartes, Chalmers, and many contemporary philosophers who find physicalist accounts of phenomenal consciousness inadequate.

Substance vs. Property Dualism

Substance dualism (Descartes) holds that mind and body are two different *kinds of stuff*, res cogitans and res extensa, leaving the interaction problem that has dogged the view ever since. Almost no contemporary philosopher defends it.

Property dualism is the live form. It grants one kind of substance (the physical) but holds that conscious experience involves a further, irreducible *kind of property* — phenomenal properties not entailed by, or reducible to, the physical facts. Chalmers's "naturalistic dualism" is the sophisticated version: one world, physically causally structured, but with phenomenal properties added as further fundamental features, connected to the physical by psycho-physical laws. See *Physicalism*.

The Case for Dualism

Dualism draws strength from a convergence of genuinely powerful intuitions, each of which feels like a discovery: the conceivability of philosophical zombies, the apparent gap in Mary's knowledge, the explanatory gap between physical description and felt quality. Each is real as an epistemic phenomenon. The dualist inference reads these epistemic facts as ontological ones: because we cannot *see* how the physical explains the phenomenal, the phenomenal must be something *over and above* the physical. This inference from explanatory gap to ontological gap is the load-bearing move in every major contemporary dualist argument. See *Zombies and Conceivability*, *Mary's Room*, and *The Explanatory Gap*.

The Standard Pressure: Mental Causation

The best-known objection to property dualism runs through causation. If the physical world is causally closed, so that every physical event has a sufficient physical cause, then irreducible phenomenal properties risk having no work to do — a consciousness that watches but never steers (epiphenomenalism). Property dualists respond by accepting epiphenomenalism, by denying closure, or by reworking the causal story; each option carries a known cost, and Chalmers, to his credit, has explored them all candidly.

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, in philosophy of mind, names the ways a system's bodily organization and environmental engagement shape perception, action, affect, autonomy, and welfare. Current internal structure alone does not fix every mental fact: causal history and relations to an environment matter. But embodiment is not one all-purpose gate through which every kind of content, reference, or consciousness must pass. Its philosophical work has to be stated capacity by capacity.

The embodiment thesis has roots in the phenomenological tradition (Merleau-Ponty's *Phenomenology of Perception*, 1945) and has been developed in various forms by cognitive scientists (Varela, Thompson, and Rosch's *The Embodied Mind*, 1991; Clark's extended-mind work) and analytic philosophers of mind (Tye's historical thesis; Burge's externalist account of objectification). Its

best use is targeted: it challenges accounts that ignore history, mode, bodily regulation, or environmental coupling, while leaving open nonbodily routes to some forms of representation.

The Brain-in-a-Vat Insufficiency

The scenario: your brain is removed from your body and sustained in a vat of nutrients, with electrodes feeding it precisely the sensory inputs it would receive if it were embodied. Functionally, it behaves identically to an embodied brain.

The classical functionalist conclusion: it has exactly the same mental states, including the same perceptual experiences.

The historical-externalist objection: the vat brain lacks the causal-historical relation that would fix its states on ordinary embodied objects. When the embodied brain represents red, its state bears a tracking relation, established over a developmental history, to actual red things in the world. The vat brain's intrinsically similar state does not inherit that reference merely by resemblance. Its actual history may instead fix content about features of the feed or a virtual environment; where no use or selection distinguishes the rivals, content may remain indeterminate.

This isn't a claim about phenomenal *feel* in isolation; it's a claim about intentional *content*. And if phenomenal character arises from representational content, the phenomenology follows.

Tye's Historical Thesis

In "Homunculi Heads and Silicon Chips" (2011), Tye argues that **history matters crucially to phenomenology**. Two individuals intrinsically identical at a time can differ in phenomenal character because of what was going on in them *in the past*.

The argument proceeds through Block's China-body system (7 billion people connected by two-tin-can phones, replicating the causal organization of a brain). Block concludes this system has no phenomenal states. Tye agrees — but argues the reason is *historical*, not organizational. The China-body system came into existence as a replication of human neural organization without the developmental history through which that organization acquired world-tracking content.

The historical thesis: what a state represents, and thereby what phenomenal character it has, depends partly on the causal history through which the system acquired that state's functional role. Content doesn't supervene solely on current intrinsic properties.

This directly parallels semantic externalism in philosophy of language: Putnam's Twin Earth and Burge's arthritis case establish that meaning depends on environment and community, not just on what's "in the head." Tye extends this to phenomenology: phenomenal character, too, depends on what's outside the head — specifically, on the causal-historical relation between the system and its environment.

What "Embodiment" Actually Requires

The embodiment thesis, in its disciplined form, resists inflating embodiment into a mystical requirement. The claim is not:

- That only biological organisms can think
- That the body as such is essential (an uploaded mind isn't ruled out in principle)

- That consciousness requires a particular physical form factor
- That every kind of content or distal reference requires direct sensorimotor contact

The positive claims divide by capacity:

1. **Perceptual and bodily-affective mode**: sensor-actuator coupling and bodily regulation help determine how a world is perceptually and affectively presented, not merely which proposition is encoded.
2. **Autonomous agency**: ongoing environmental coupling, self-regulation, and action-feedback loops can support a system that initiates and corrects activity across time.
3. **Stake and welfare**: self-maintaining organization can give outcomes practical significance for the system — something can go well or badly for it.
4. **Content and reference**: causal tracking history and producer-consumer use help fix what states represent, but direct bodily contact is only one route. Social inheritance and corpus-mediated causal chains may transmit public reference; training can make derived content causally internal to a mechanism. Original content remains possible where a world-involving history fixes correctness independently of an interpreting practice.

A sufficiently embedded robotic system could realize all four together. A text-only model may realize parts of the fourth without realizing the first three. That is a more discriminating verdict than either "text is enough for a mind" or "no body, no content."

Semantic Externalism as Background

The embodiment argument inherits from semantic externalism:

- **Putnam (1975)**: "meanings just ain't in the head." Twin Earth cases establish that what a term refers to depends on the causal-environmental history of its use, not on internal mental states.
- **Burge (1979)**: what concepts a person has depends partly on their social and linguistic environment, not purely on their internal organization.
- **Kripke (1980)**: reference is fixed by causal-historical chains linking uses back to original baptisms, not by descriptive content in the speaker's head.

The embodiment thesis adds a realization story for some externalist relations: bodies make environmental tracking, correction, action, and regulation continuous and consequential. But linguistic practices can also transmit reference through social and causal-historical chains, and training can make derived linguistic content causally operative inside learned producer-consumer mechanisms. A purely formal description supplies no semantics; it does not follow that every text-trained mechanism is causally or semantically isolated.

The Inner Monologue Connection

On a Vygotskian line, inner speech is internalized social behavior — linguistic activity that began as public, world-involving exchange and was gradually internalized. This reinforces the historical and social point: even apparently inner activity may inherit structure and content from public practice. It does not establish that each user, or each mechanism trained on the record of that practice, must reenact the original bodily encounters to participate semantically.

Implications for AI

Embodiment matters directly to AI where the claimed capacity depends on perceptual mode, bodily affect, autonomous action, self-maintenance, or welfare. Current text-only LLMs lack the continuous sensorimotor and regulatory organization that would make those claims straightforward. Yet training does give them a causal history: their corpora descend from human world-engagement, and selection shapes internal mechanisms for use by downstream processes. That history may establish content without establishing perception, feeling, autonomous agency, or whole-system understanding. Future architectures with ongoing environmental feedback would strengthen several of these dimensions at once, but the dimensions should not be collapsed.

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 identifies mental states by their causal-functional roles: what they receive as input, what they produce as output, and how they relate to other mental states. Pain, on this view, doesn't name a particular neural state; it names the functional role of being caused by tissue damage, producing avoidance behavior, and interacting with beliefs and desires in characteristic ways. Silicon or carbon, neurons or circuits — whatever plays that role is in pain.

Functionalism emerged in the 1960s as a response to the perceived failures of behaviorism (which eliminated inner states) and identity theory (which tied mental states too tightly to particular physical substrates). Hilary Putnam's multiple realizability argument provided the founding case: if the same mental state can be realized by different physical systems, then mental states cannot be identical to neural state types. This opened the door to **computational functionalism**, a more specific theory that identifies the relevant functional roles with computational states. Functionalism itself does not say that minds are programs.

The Core Thesis

Putnam's original formulation (1960s): mental states are multiply realizable. The same mental-state type can be realized by different physical substrates across different organisms and possible

systems. A Martian can share our mental states without sharing our neurology, provided its internal organization plays the same functional role.

The functionalist move: define mental states by their input-output profile and their relations to other mental states — not by what they're made of. That move leaves open how the roles should be specified. **Computational functionalism** specifies them as states in a machine table or program. The hardware-software picture belongs to that child theory, not to functionalism as such.

What draws philosophers to it:

- Mental states have causal powers. Beliefs and desires don't just accompany behavior; they produce it in systematic ways.
- Multiple realizability commands wide assent — there is no obvious reason only neurons could think.
- The organizational structure of a system matters for its mental properties, not just the raw physical substrate.

Block's Objections

Ned Block's "Troubles with Functionalism" (1978) remains the most influential critique. Two objections define the dialectic:

1. The Absent Qualia Problem

Imagine a system that plays the right functional role (same input-output profile, same internal state transitions) but has no phenomenal experience at all. Block's **homunculus-head**: replace each neuron in your brain with a tiny person communicating by two-tin-can telephone. The system as a whole plays the functional role of a human mind. Does it feel anything?

The intuition: probably not. But functionalism cannot say why not — by its own criterion, the homunculus-head qualifies as having mental states.

The book's diagnosis: the objection bites against narrow, world-bracketed functionalism. An internal diagram can reproduce a pattern of transitions without supplying the history that fixes what its states represent. Even content leaves further questions: which states are conscious, which persisting organization bears them, and what complete world-presenting profile determines the character of states already counted conscious. A system can play a stipulated pain-role without answering any of those questions.

2. The Inverted Qualia Problem

Imagine two people with systematically inverted color experience: what produces "red experience" in you produces "green experience" in your twin, but your behaviors are identical throughout — you both call grass "green" and stop at red lights. Functionally, you're identical. Phenomenally, you're inverted.

If functionalism is true, you have the same mental states. But the phenomenal character differs. So phenomenal character outruns functional role.

On the representationalist reading, this argument works against bare internal role as a complete theory of phenomenal consciousness, but it does not establish that qualia are non-physical.

Content-fixing history, phenomenal mode and profile, the conscious threshold, and subject individuation require separate specifications rather than one enlarged functional role.

Tye's Response: Functionalism + Content

Tye's move (in *Ten Problems of Consciousness* and "Absent Qualia and the Mind-Body Problem"): functionalism fails not because phenomenal character is non-functional, but because it left out *representational content*. The right theory combines functional role with what the state *represents*.

On this account, Block's stipulated narrow diagram does not establish phenomenal consciousness. A complete assessment must ask what its states represent, whether any enter consciousness, which continuing control organization bears them, and what world-presenting profile conscious states instantiate. A homunculus-head that genuinely satisfied those conditions would qualify; bare transition-role equivalence settles none of them.

Tye's historical **PANIC theory** (Poised, Abstract, Nonconceptual, Intentional Content) supplies one influential version of the representational repair. The book retains the identity claim while treating PANIC as one proposal about perceptual form and the conscious threshold, not as a complete account of content or the subject.

The Chinese Room and the Sufficiency Claim

Searle's Chinese Room targets the computational version of functionalism directly, but its clean result is narrower than Searle's verdict. The operator's ignorance shows that rule following by the implementation substrate does not establish operator-level understanding; fluent output does not establish understanding by the implemented whole either. Whether a sufficiently rich virtual organization understands must be judged from its content-fixing relations and coordinated capacities, not settled by the operator's introspection.

The systems reply correctly relocates the question: perhaps the whole organization understands even though the operator does not. Internalizing the rulebook sharpens the case but does not decide that whole-system attribution. The Room supplies public Chinese content and formal processing; it does not establish that one persisting system integrates those contents across perception, memory, correction, inference, planning, and action. The robot and combination replies sketch how a machine might add the missing organization. The Room therefore defeats syntax-alone sufficiency, not artificial understanding.

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, practical interest, and welfare. It neither creates content nor constitutes understanding or consciousness. No part of this account privileges neurons, carbon, or biological manufacture.

Higher-Order Thought (HOT) Theories

Higher-order thought theories explain what makes a mental state *conscious* by appeal to higher-order mental states — thoughts or perceptions directed at the first-order state itself. A pain is conscious not because of anything intrinsic to the pain, but because one has a thought (or perception) *about* the pain. Unconscious mental states are states one is not in any way conscious of being in.

The view originates with David Rosenthal (City University of New York), who developed it as a way to explain the commonsense asymmetry between conscious and unconscious mental states while remaining within a broadly physicalist framework. HOT theories represent the most developed alternative to intentionalism for explaining phenomenal consciousness without invoking qualia as primitive inner properties.

The Transitivity Principle

Rosenthal's foundational insight. When we speak of a mental state as "conscious," we mean something close to what Descartes and Locke meant when they said one is *conscious of* one's mental states. A state is conscious if, and only if, one is conscious of being in it in some suitable way.

The Transitivity Principle (TP): a mental state S is conscious iff one is transitively conscious of S.

This is not circular: the higher-order state that makes S conscious need not itself be conscious. A subliminal perception makes one conscious of an object without being conscious itself. Equally, a HOT that makes a pain conscious need not be a conscious thought — and usually isn't.

TP receives support from the direction of explanation: no state seems conscious if one is in no way whatever aware of being in it. The question is only *how* we are conscious of our conscious states.

Two Versions

1. Inner Sense (Higher-Order Perception)

Armstrong and Lycan: a mental state is conscious when a quasi-perceptual inner monitoring mechanism tracks it. Consciousness involves a kind of inner sense directed at first-order states. The monitor relays information about ongoing psychological events, enhancing rationality and coordination.

Rosenthal's objection: when we are conscious of our mental states, no mental quality belongs to the higher-order awareness itself — the awareness of pain has no intrinsic quality of its own. But perceiving always involves some qualitative character. Therefore our higher-order awareness of mental states cannot be perceptual. It must be cognitive — a *thought* about the state.

2. Higher-Order Thoughts (Rosenthal)

The canonical version. A mental state is conscious in virtue of being accompanied by a suitable HOT — an assertoric, non-inferential thought to the effect that one is in that state right now.

Key features: - **Non-inferential**: the HOT must not be mediated by inference; otherwise the resulting consciousness would feel mediated, which it doesn't - **Assertoric**: the HOT asserts rather than doubts or wonders; doubting doesn't make one conscious of anything - **Occurrent, not merely dispositional**: being *disposed* to have a HOT doesn't make one conscious of a state; one must actually have the HOT (contra Carruthers' dispositional variant) - **The HOT itself rarely conscious**: HOTs are conscious only when accompanied by third-order thoughts — which seldom happens. This is why we don't notice being in two states at once when we have a conscious experience

Evidence Rosenthal offers: - *Reportability*: we determine whether a state is conscious by whether one can report it. Reports express thoughts about one's states. So conscious states are reliably accompanied by thoughts about them. - *Fine-grained variability*: learning new words for qualitative properties sometimes makes one conscious of experiences in more fine-grained ways. This suggests HOTs (which employ concepts) govern the character of conscious experience. - *Dental fear*: patients with anaesthetized nerves sometimes feel pain due to fear and the sensation of the drill's vibration. The HOT (about being in pain) determines what it's like for them, not the first-order state.

Carruthers' Dispositional Variant
Carruthers objects that Rosenthal's view posits two states when experience seems to involve only one. He proposes: a state is conscious if a HOT is *disposed* to occur, not if one actually occurs.

Rosenthal's reply: dispositions cannot implement TP. Being disposed to be conscious of something doesn't make one conscious of it. And the dispositional theory cannot explain introspection as anything more than a disposition-to-a-disposition, losing explanatory traction.

Carruthers' later move: build higher-order content *intrinsically* into the first-order state (via a disposition-to-cause-a-HOT that confers higher-order content). The conscious state itself has the content that it represents itself as a mental state. This eliminates the two-state worry. Kriegel (self-representationalism) develops a version of this independently.

The Core Objection: HOTs Without Targets
If a HOT can occur without a suitable first-order state, there should be a phenomenally conscious experience with nothing "underneath" it. Rosenthal bites this bullet: the HOT alone would suffice for what it's like for one to have that experience. The dental fear case supports this — what it's like is determined by how one is conscious of a state, not by the state's intrinsic properties.

The mismatch problem: if the HOT misrepresents the first-order state (HOT says "I'm in pain" but the first-order state is fear), what is the experience like? On TP, it's like being in pain — because what it's like is determined by the HOT's content, not the state's intrinsic properties.

This generates a genuine puzzle for HOT theories about the relationship between phenomenal character and neural reality.

What HOT Theories Get Right
The view has genuine strengths:

- **The distinction between conscious and unconscious mental states**: there clearly are unconscious mental states (subliminal perception, masked priming), and HOT theories offer a principled explanation of what makes the difference
- **The role of self-awareness in consciousness**: there is something right about the idea that conscious experience involves a kind of awareness of one's own states — the question is whether this is higher-order in Rosenthal's sense or whether transparency provides a different explanation
- **The reportability connection**: the link between consciousness and verbal report is real; HOT theories explain it neatly
- **The empirical orientation**: Rosenthal takes cognitive science seriously and tries to make HOT theories testable — a genuine virtue

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 (Mind-Brain Identity Theory)

Identity theory holds that mental states just *are* brain states — not correlated with them, not caused by them, not realized by them in some softer sense, but numerically identical with them. When you have a pain, the pain is identical to a particular neural process. When you have an after-image, the experience is identical to a brain state. One thing, described two ways.

It was the first fully serious physicalist theory of mind, and it remains the pivot point in the history of the field: it marks the moment philosophy of mind committed to physicalism, and the pressure it came under directly produced functionalism.

Place and Smart

U. T. Place, "Is Consciousness a Brain Process?" (1956) — the opening shot. Place argued that consciousness, like lightning or genes, is a phenomenon that turns out upon scientific investigation to be a physical process. The identification is an empirical discovery, not a conceptual analysis. Just as lightning *is* an electrical discharge (not merely correlated with one), consciousness *is* a brain process.

J. J. C. Smart, "Sensations and Brain Processes" (1959) — the canonical elaboration. Smart's motivation: physicalism about biology, chemistry, and behavior is increasingly plausible, leaving consciousness as the sole "nomological dangler" — a phenomenon that hangs outside the physical web of laws by a mysterious thread. The clean solution: sensations *are* brain processes.

Smart's key move: the identity is *contingent*, not definitional. "After-image" and "brain process of sort X" don't mean the same thing — just as "lightning" and "electrical discharge" don't mean the same thing. An ancient Greek peasant could correctly report his after-images without knowing anything about neuroscience, just as he could correctly identify lightning without knowing anything about electricity. Meaning doesn't determine reference.

The **topic-neutral language** strategy handles the trickiest objection. When someone says "I have a yellowish-orange after-image," Smart analyzes this as: "there is something going on which is like what is going on when I have my eyes open, am awake, and there is an orange illuminated in front of me." The phrase "there is something going on which is like" is topic-neutral: it doesn't specify whether the something is physical or phenomenal. It leaves open what the something is, which is exactly what's needed for a contingent identity claim.

The nomological danglers argument: if sensations were something over and above brain processes, there would have to be psycho-physical laws connecting them. But these laws would be

irreducibly queer — laws connecting enormously complex neural configurations to non-physical somethings, unlike anything else in science. Parsimony and simplicity favor the identity theory: eliminate the danglers by identifying the mental with the physical.

The Historical Position

Identity theory sits between behaviorism and functionalism in the dialectical sequence:

Behaviorism (Ryle, early Wittgenstein): mental states are not inner things at all — they are behavioral dispositions or patterns. "He is in pain" means roughly "he is disposed to wince, cry out, reach for analgesics." No inner states, no privacy, no problematic non-physical entities. Clean and scientifically respectable.

Problem with behaviorism: the field came to regard it as plainly false. Pain is not a disposition to behave in certain ways — it's something that *causes* that behavior, and it has an intrinsic qualitative character that no behavioral disposition captures. You can have pain without any behavioral manifestation (stoics exist); you can have all the behavior without pain (actors, robots).

Identity theory accepts that there are genuine inner states (reports of sensations are genuine reports, not mere behavior) but identifies those inner states with brain states. Physics gets the full explanatory load; no separate domain of the mental is needed.

Problem with identity theory: it requires that the same mental type is always realized by the same neural type. But pain in a human, pain in an octopus, and pain in a silicon robot might be realized by completely different physical structures. Multiple realizability, the same mental state type realized by different physical substrates, appears to refute the type identity claim. The neural state type in humans can't be identical to the neural state type in octopuses if octopuses have completely different neurons. Yet they might both have pain.

Functionalism emerges from this pressure: instead of identifying mental types with physical types, identify them with functional roles — causal-organizational structures that can be realized by multiple physical substrates. This preserves the physicalist spirit of identity theory while accommodating multiple realizability.

The Standard Objections

Smart systematically addressed eight objections in the 1959 paper. The most important:

The knowledge objection (Objection 1): an illiterate peasant knows what after-images are without knowing anything about brain processes. Therefore after-images cannot be brain processes.

Reply: the Morning Star / Evening Star case. An English sleeper who never sees the Morning Star still refers to the same thing as the early riser who uses "the Morning Star." Ignorance of the identity doesn't prevent reference. The peasant reports his after-image as "something going on like what goes on when I see an orange" — topic-neutral, leaving open what the something is.

The qualitative properties objection (Objection 3, which Smart regarded as hardest): even if we identify experiences with brain processes, there are irreducibly psychic *properties* (being yellowy-orange, being painful) that brain processes don't have. The identification of states leaves out the qualitative character.

Reply: Smart's topic-neutral strategy again. Reporting a yellowish-orange after-image is reporting something that is like what goes on in normal orange-perception. No irreducibly phenomenal predicate appears in the report. The qualitative character is described by resemblance to typical perceptual situations, not by intrinsic phenomenal properties.

The after-image in space objection (Objection 4): the after-image isn't in physical space; the brain process is; so they can't be identical.

Reply: the claim is not that the after-image is a brain process; the *experience of having* an after-image is. The experience of seeing yellowy-orange is not itself yellowy-orange. The surgeon looking into the brain won't see a yellowy-orange thing because the experience is not a yellowy-orange thing — it's a brain state that represents yellowy-orange things.

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 perception never makes direct contact with the mind-independent world. What we perceive immediately and directly are inner objects (sense-data, phenomenal intermediaries, the "veil of perception"), and the external world is known only by inference from these inner items.

The view has deep historical roots. Its lineage runs from Locke's primary/secondary quality distinction through Berkeley's idealism, Hume, the British Empiricists, through Russell's early work, and through the sense-datum theorists of the early twentieth century (Moore, Price, Ayer). Its central argument, the argument from illusion, remains the most discussed argument in the philosophy of perception. Understanding indirect realism precisely matters because it is not obviously wrong; the argument for it is genuinely compelling, and the history of philosophy of perception is largely the history of attempts to answer it.

The Argument from Illusion

The engine that drives people into indirect realism. Three premises:

1. **The perceptual variation premise**: the same external object can produce different perceptual experiences in different conditions (lighting, angle, medium, state of the perceiver). A

round coin looks elliptical from an angle; a white shirt looks blue in shadow; a straight stick looks bent in water.

2. **The phenomenal object premise**: in these cases, there *is* something that has the property the experience presents — something that genuinely is elliptical, blue, bent. This something cannot be the external object (the coin is round, the shirt is white, the stick is straight). So it must be an inner object: a sense-datum.
3. **The parity premise**: since veridical and illusory experiences can be subjectively indistinguishable, the immediate object of perception in *both* cases must be the same kind of thing — a sense-datum. In veridical perception, we see sense-data that accurately represent the world; in illusion, sense-data that misrepresent it.

Conclusion: we never perceive the external world directly. We perceive sense-data directly; the world only mediately, by inference.

Berkeley pushes the same lineage one step further — strip out the external world entirely, and only ideas remain.

What Sense-Data Are Supposed to Be

On the standard sense-datum theory:

- **Private**: sense-data exist in the perceiver's mind, not in the public world
- **Directly apprehended**: we are infallible about their intrinsic properties — if the sense-datum looks elliptical, it *is* elliptical
- **Phenomenally characterized**: they possess the sensible qualities experience presents (color, shape, texture) as intrinsic properties
- **Causally produced by but numerically distinct from external objects**: the round coin causes an elliptical sense-datum; the external object and the inner object come apart

The inner theater follows immediately from this picture: the mind becomes a private theater in which sense-data perform, and the perceiver is an audience of one, watching an inner show that represents the external world but never is it.

Sellars on the Myth of the Given

Sellars's "Empiricism and the Philosophy of Mind" (1956) provides the foundational analytic attack on the sense-datum picture, prior to and more general than the contemporary intentionalist critique. The *Myth of the Given* is the view that empirical knowledge rests on epistemically self-authenticating immediate experiences — items that are simply *given* to consciousness, presenting themselves directly and incorrigibly, and serving as the foundational evidence base for all further knowledge claims.

Sellars's attack proceeds in two stages.

First, the structural diagnosis: the Given must do two incompatible jobs. It must be (a) immediate, non-conceptual, available without prior knowledge, or it is not a foundation; and (b) the kind of thing that can serve as a reason for empirical beliefs, or it cannot ground knowledge. But nothing can be both. To serve as a reason, an item must enter the logical space of concepts — must be the

kind of thing that can stand in inferential relations to beliefs. And anything that does that has already been conceptualized. The supposedly *raw* given is a fiction.

Second, the Jonesean myth offers the positive reconstruction. Thought-talk and sensation-talk are not reports of pre-given inner items but theoretical posits modeled on the antecedent practice of public discourse. We do not first encounter inner sensations and then learn to describe them; we first learn the public language of perception (talk of red surfaces, of seeing things correctly or incorrectly) and then learn, via something like Jones's theoretical move, to apply that vocabulary to inner episodes. The order of explanation runs from world-directed perceptual talk to talk of inner states, not the other way.

The bearing on sense-datum theory is direct: the sense-datum theorist supposes that the inner colored patch (or inner phenomenal item) is the foundational given, the thing perception immediately delivers, prior to all theory. Sellars shows this is incoherent. Whatever the senses deliver, they do not deliver self-authenticating inner objects ready to ground inference. The picture of the inner colored patch as immediate datum is itself a piece of theory — a theory that, on examination, conflates the (legitimate) thought that perception has phenomenal character with the (illegitimate) thought that phenomenal character consists in immediately-presented inner objects.

McDowell's *Mind and World* (1996) develops the Sellarsian line into its principal contemporary form: rejection of self-standing inner standpoints, second-nature realism, the unified spontaneity/receptivity picture. The intentionalist case against sense-data inherits Sellars's structural diagnosis and McDowell's contemporary deployment.

The Neuroscience Version

The sense-datum picture survived into contemporary neuroscience under different vocabulary. The brain-as-screen model: perception involves the brain constructing an internal model or representation of the world, and what consciousness has access to is this model — not the world itself. Sometimes it travels as "controlled hallucination" (Anil Seth); Dennett attacks a version of it as the Cartesian Theater. The basic structure is unchanged: inner representational objects stand between the perceiver and reality.

The philosophical objection to this version: it runs together (a) the neural processes that *realize* perception and (b) the objects that perception is *about*. Describing how the brain produces perceptual states doesn't establish that what we perceive is a brain state rather than the world.

The Transparency Diagnosis

The central intentionalist move against indirect realism: **the argument from illusion rests on the phenomenal object premise, and that premise is false**.

When I see the bent-looking stick, there is indeed an experience with a particular character. But that character consists in *how* the stick is represented, not in the properties of an inner object. The experience represents the stick as bent — that's what it is for the experience to have bent-stick character. No inner bent thing exists; there is only an outer (straight) thing represented incorrectly.

The Transparency Thesis dissolves the argument from illusion by denying that introspection reveals a second inner object. When you attend carefully to your visual experience of the red apple, awareness returns to the apple — its redness, roundness, and position — while also allowing you to notice the visual mode in which those features are presented. That mode belongs to the complete intentional structure of the experience; it is not an inner colored surface or non-intentional residue.

On this diagnosis, **the phenomenal object premise** mistakes the representational character of experience (how things are presented) for properties of an inner object (what things are present to awareness). Reject that premise, and the argument from illusion doesn't establish sense-data — it establishes only that perceptual experiences can misrepresent their objects. Misrepresentation requires no inner objects; it requires only that a state with a certain representational content fails to match the world it represents.

Adverbial Theory
An intermediate attempt to preserve the argument's grip while avoiding inner objects. Instead of saying "S perceives a red sense-datum," say "S perceives red-ly" or "S senses red-ly." The adverb modifies the perceptual act, not a perceived object.

The many-property problem (Jackson): if I simultaneously experience a red triangle and a green circle, the adverbial theory must distinguish "experiencing red-triangularly and green-circularly" from "experiencing red-circularly and green-triangularly." The object-theory handles this by individuating the objects; the adverbial theory must introduce complex adverbial modifiers that replicate object-theory structure without the objects. The simplicity gain disappears.

After Jackson's objection, few remained adverbialists; in the field's retrospect the view reads as a way-station, not a destination.

Disjunctivism
The direct-realist alternative that acknowledges the parity premise's bite while rejecting its conclusion. McDowell, Hinton, Snowdon, Martin:

The disjunctivist move: veridical perception and subjectively indistinguishable hallucination are not instances of the same fundamental kind of experience. In veridical perception, the subject stands in a direct relation to the mind-independent object — the object *itself* is a constituent of the experience. In hallucination, no such relation obtains; the experience has only an internal cause. These are disjuncts, not instances of a common kind.

The parity premise fails: the subjective indistinguishability of perception and hallucination doesn't establish that they share the same immediate object. Two situations can be indistinguishable from the inside without being fundamentally the same type of situation.

A further pressure case, due to Tye: a hand hidden behind a screen causes, by a deviant route, a visual experience as of a hand. Naive realism struggles to say why this is not perception *of* that hand; intentionalists answer such cases by specifying the right kind of causal-representational relation. Deviant causal chains remain a standing test for both camps.

The Inner Theater Picture

What indirect realism entrenches, its critics charge, is a whole picture of mind:

- **Mind as container**: the mind is a space inside which mental objects reside
- **Experience as object**: experiences are things we encounter, not relations we stand in
- **Perception as two-stage**: first inner object, then inference to outer world
- **Privileged access as infallibility**: we cannot be wrong about inner objects because they are constituted by how they appear

On the critics' diagnosis, each element generates philosophical problems that wouldn't exist without the picture: skepticism about the external world, the problem of other minds, the problem of phenomenal properties. All presuppose that inner objects exist and must somehow bridge to a world they stand apart from.

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" name different achievements. A system can carry and process information without every informative state having semantic content. The first question is whether anything fixes an accuracy condition at all. A further question then asks whether that content is derived from an interpreting practice or original because its correctness standard is fixed independently of one. Ownership settles neither question.

Two Senses of Information

Shannon information is a measurable statistical relation: one state carries information about another when it reduces uncertainty about it. Smoke carries information about fire; tree rings about age; a thermometer's strip about temperature. This notion by itself contains no truth, reference, or error.

Semantic content concerns what a state represents and how it can misrepresent. A representation can be causally effective while false. That possibility requires more than covariance.

Dretske's Bridge

Fred Dretske tried to naturalize content by beginning with lawful indication and adding a history in which an organism learns to use the indicator. The acquired function distinguishes what the state was recruited to indicate from whatever happens to cause it now. Teleosemantics generalizes the strategy through selected producer-consumer organization.

The bridge succeeds only if the proposed mapping does genuine explanatory work. It should help explain why the mechanism was stabilized, predict consumer success and failure, and remain privileged under interventions that separate rival correlations. When several mappings remain explanatorily equivalent, content may be indeterminate at that grain. When the mapping merely repeats an engineer's intention, the content remains derived.

Observer Dependence

Searle is right that unrestricted "information processing" talk can be observer-relative. Almost any physical system permits some mapping from its states to symbols. Calling one mapping semantic explains nothing unless the mechanism itself supplies constraints that privilege it.

But it does not follow that physical mechanisms contain only causal traffic until an owner appears. Selection, learning, and artificial training can organize producers and consumers so that one correspondence rule explains performance better than its rivals. Such a state may have non-derived mechanism-level content. The state need not be a subject, and its containing system need not understand it.

The AI Case

A text-trained language model contains causally potent states sensitive to linguistic distributions. If a linguistic-context mapping survives interventions on training, producer states, downstream consumers, and loss-relevant success, those states may carry causally internalized derived content at that narrow level. The vocabulary, corpus, task, and correctness standard still descend from public linguistic practice. This grants more than the claim that an engineer's interpretation is non-arbitrary without pretending that training originated the semantic standard.

The grant remains limited. Change the world while holding the corpus fixed and the trained mechanism does not follow the world; change the corpus while holding the world fixed and it follows the corpus. Textual content need not amount to reference to particular distal objects. Nor does content in a submechanism establish whole-model belief, understanding, perception, or consciousness.

Those larger attributions depend on how content-bearing states are integrated and poised across memory, learning, inference, conflict resolution, planning, and action. Self-maintenance may bear on whether the system counts as an autonomous agent. It does not turn information into 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 the phenomenologically vivid silent verbal activity that pervades ordinary conscious thinking — the "inner voice" people report when they rehearse a remark, work a problem under their breath, argue with themselves, or read a sentence to themselves. It feels like the most private region of the mind: a first-person stream, audible only to its owner, the deepest interior we have.

Research on inner speech spans psychology, linguistics, and philosophy of mind. Empirical work (Alderson-Day and Fernyhough's survey literature, Vygotsky's developmental account) documents its pervasive role in reasoning, memory, and self-regulation. Philosophical work focuses on its relation to linguistic meaning, semantic externalism, and the nature of conscious thought.

Two Pictures of the Inner Voice

The standard picture runs: inner thought comes first, and public speech externalizes it. The rival picture, descending from Vygotsky, inverts the order: public speech comes first, and the inner voice is that speech internalized — the same activity, run silently. For critics of the standard picture, a great deal hangs on the choice. Once the inner voice counts as a private inner thing, an interior space with its own contents, accessible only to its owner, has already been conceded — the Cartesian theater, admitted through the side door.

The inversion is not the claim that all cognition is inner monologue. Pre-linguistic infants think; non-human animals think; expert performance often runs faster than any inner narrator could keep up with. The internalization claim is narrower: *the experience of thinking-in-words*, the phenomenologically vivid inner voice, is internalized public speech. Non-verbal cognition may be conscious or unconscious and may have sensory, affective, or other phenomenal character; it simply is not the verbal phenomenon under discussion here.

The Two Anchors

Two anchors hold the position in the literature.

The Wittgensteinian anchor. The private-language argument (*Philosophical Investigations* secs. 243–315, especially sec. 258) argues that meaning cannot be fixed by what goes on inside a single skull. Following a rule requires the possibility of public correction; without that, there is no fact of the matter about whether the next use of a term agrees with the prior ones. Inner-voice meaning,

if it floated free of any corrective practice, would not be meaning at all. The argument generalizes: meaning gets fixed in the 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.

The Sellarsian anchor. "Empiricism and the Philosophy of Mind" (1956), especially the Jonesean myth in secs. 48ff., provides the positive picture. In Jones's story, thought-talk is theoretical posit, *modeled on the antecedent practice of public discourse*, not discovered by inward inspection of pre-given inner items. Thought-talk is conceptually downstream from speech-talk, even when its referents are silent. We learn what thinking is by learning to talk; then we learn to call the silent version of that activity "thinking."

Crane and Farkas (2022) crystallize the picture in one sentence: thoughts are "inner episodes, called 'thoughts,' which are conceived on the model of overt verbal utterances, but happen silently in the head." Their own positive thesis applies the modeled-rather-than-inspected picture to standing mental states (beliefs, desires); the internalized-speech account extends it to occurrent inner speech episodes.

What Falls Out

If the internalized-speech account is right, three consequences follow.

(1) Meaning is borrowed, not generated. The inner voice uses words you did not invent, in a language you did not design, calibrated by a community you mostly never met. This is what semantic externalism looks like once you take it seriously about silent thinking, not only about overt linguistic reference. The Putnam/Burge externalist line about meaning extends to inner speech without modification.

(2) The inner voice is not the inner theater. Once inner speech is public speech gone quiet, the apparent privacy of inner monologue dissolves into the linguistic competence that produced the public speech in the first place. The "inmost private chamber" is the public speaker, rehearsing in a register quiet enough that only the speaker hears.

(3) LLM-style inner-voice-like output does not by itself establish an inner life. Fluent, corrigible language can provide defeasible evidence of content-guided competence, and training can make inherited public meanings causally operative inside a mechanism. Model outputs may also refer through the linguistic and social practices from which they descend. None of that alone establishes a persisting system in which memory, correction, planning, perception, and action constrain one another, still less a perceptual or bodily-affective mode poised for consciousness. The output is evidence to be weighed with provenance and architecture, not proof either of an inner speaker or of its absence.

Inner Speech vs. Cognitive Phenomenology

The cognitive-phenomenology literature (Pitt, Strawson, Siewert, Bayne and Montague's volume) presses the claim that occurrent conscious thought has a proprietary phenomenal character distinct from sensory imagery. The internalized-speech account is compatible with this claim about *what* the phenomenology is like, while differing on its explanatory direction: the linguistic

character of inner thought episodes is not evidence of a private language but of the public language doing its work silently.

Jorba and Vicente (2014) argue specifically that cognitive-phenomenology defenders are best positioned to explain why inner speech is so pervasive in conscious thinking. On the internalized-speech reading, that diagnosis holds, but pervasiveness is itself a clue to provenance: inner speech is everywhere in conscious cognition because public speech is everywhere in social life, and the silent register inherits whatever the loud register supplied.

Inner Speech vs. Language of Thought

Fodor's Language of Thought hypothesis posits a syntactic medium of cognition (Mentalese) underlying both public speech and inner speech. The internalized-speech account differs: inner speech is not the *outer surface* of a deeper inner code. It is itself the public language, run quietly. The Mentalese hypothesis treats public language as expression-of-thought; the inner-speech-as-internalized-public-speech picture treats public language as the constitutive medium that thought, in its phenomenologically vivid form, inherits.

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

What Intentionalism Claims

Intentionalism is the thesis that the phenomenal character of experience, what it is like to undergo it, consists in, or at least supervenes on, its representational content. Two experiences that represent the world in exactly the same way cannot differ in what it is like to have them.

Weak intentionalism (supervenience): Necessarily, experiences alike in representational content are alike in phenomenal character. This is silent on the nature of phenomenal character — it just says content fixes phenomenology.

Strong intentionalism (identity): Phenomenal character *consists in* a certain kind of representational content. There is no residue, no additional inner property over and above the content. Tye and Dretske supplied influential contemporary versions; Harman, Moore, Crane, Block, Byrne and Hilbert, Millikan, and others supply evidence and constraints. Strong intentionalism and **strong representationalism** name the same identity claim. The book adopts it as a cumulative synthesis rather than on any one figure's authority.

Why Intentionalism

The opening motivation is transparency. When you attend to your experience, you find the world as presented — surfaces, colors, shapes, sounds — and the mode, determinacy, field structure, or attention through which it appears. You do not find a second inner bearer of those qualities. That datum removes direct acquaintance with mental paint as evidence; it does not deductively exclude a hidden residue, and it does not by itself choose historically wide content over a synchronic structural presentation.

The identity claim enters as a cumulative inference to the best explanation. Independently fixed content, perceptual or bodily-affective mode, determinacy, field structure, and attention account for the phenomenal contrasts examined in the book, while no independently identified inner property predicts a contrast that complete profile misses. A residue-free grounding theory and a structural representationalism remain genuine rivals wherever they match that explanatory record. The book adopts identity because one disciplined profile carries the work with fewer commitments, not because transparency has already proved it.

A further motivation follows from naturalization. If phenomenal character consists in representational content, and representational content can receive a causal, teleological, or environmental account, consciousness fits within a physicalist picture without an ontological remainder. State consciousness and subject individuation still require separate evidence.

Tye's PANIC Theory

Tye's historical framework specifies Poised, Abstract, Nonconceptual, Intentional Content — PANIC. The book treats it as one influential proposal about perceptual form and the relation marking state consciousness, not as a complete theory of content, subjecthood, or realization.

- **Poised:** The content stands ready to be used by cognitive systems for reasoning, belief-formation, and behavioral control. It is "at the interface" between sensation and cognition.
- **Abstract:** The content abstracts away from particular instances — it represents redness, not this particular red thing.
- **Nonconceptual:** The subject need not possess the concepts for the properties represented. A creature without the concept "red" can still have an experience as of red.
- **Intentional:** The content is representational — it has accuracy conditions, it is about something.

Intentionalism vs. Qualia Realism

Qualia realists (Block, Shoemaker, Levine) hold that there are properties of experience (qualia) that are intrinsic, non-relational, and not exhausted by representational content. On this view, two creatures could represent the world identically while differing in phenomenal character (inverted qualia).

Intentionalism denies this. On the intentionalist diagnosis, the appearance of a residue beyond content results from a misleading picture: the Cartesian inner theater, in which experience presents an inner surface to an inner observer. Once that picture dissolves, via transparency, the motivation for qualia realism collapses.

Pitt's Cognitive-Phenomenology Version

A recent and substantive ally for strong intentionalism comes from a different direction. David Pitt argues (especially in "Introspection, Phenomenality, and the Availability of Intentional Content," in Bayne & Montague 2011) that the phenomenal character of an occurrent conscious thought *is* its intentional content. The phenomenology of thought is proprietary (a sui generis kind, distinct from auditory or visual phenomenology), distinctive (different thought-types have different phenomenal character), and individuative (the phenomenal character of a conscious occurrent thought is what makes it the thought it is, with the specific intentional content it has).

Pitt arrives at strong intentionalism through the cognitive-phenomenology debate rather than through transparency about perceptual experience. The two routes are complementary: Tye's perceptual route and Pitt's cognitive-thought route converge on the same identity claim from different starting points. Where Tye argues that we cannot find perceptual phenomenology apart from world-directed representational content, Pitt argues that we cannot find the phenomenology of *thinking that p* apart from the content *that p*. The two routes reinforce each other.

Strong intentionalism sides with Pitt against rival accounts of cognitive phenomenology that would treat it as additional inner-phenomenal residue floating free of content. Where Pitt differs from Tye on what kinds of content can be phenomenal, the book keeps the broader constraint: content needs independently grounded accuracy conditions, while phenomenal mode and state consciousness remain separate questions.

Intentionalism vs. Higher-Order Theories

Higher-order theories (Rosenthal) hold that phenomenal consciousness requires a higher-order representation *of* the first-order mental state. Intentionalism holds that the complete first-order structure can suffice — no meta-representation required. The debate turns on whether a state can be phenomenally conscious without being represented as such.

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 holds that perception puts us in direct contact with mind-independent objects, and that it does so *through* representational content rather than in spite of it. You see the tomato itself — red, ripe, out there on the counter — not an inner image of the tomato, not a sense-datum standing in for it. What makes that directness possible is that your experience represents the world as being a certain way, and in veridical perception it gets the world right. Representing the world does not block the world. Representing the world is what reaching it comes to.

The name pins down a specific pair of commitments. *Direct realism* fixes what perception reaches: the external object, unmediated by any inner proxy. *Intentionalist* fixes how it reaches it: in virtue of intentional content — how the experience presents things, with accuracy conditions it can meet or miss. Put together, the view occupies a narrow and much-contested piece of ground, flanked on one side by theories that insert a mental intermediary and on the other by direct realisms that refuse content altogether.

What the View Claims

Two theses, held at once.

Directness. In veridical perception we are aware of mind-independent objects and their properties, not of any mental surrogate. The apple is red whether or not anyone looks; when you look, the apple itself is what you see. No sense-datum, phenomenal image, or neural picture stands between perceiver and world as a further object of awareness.

Content. Perceptual experience has representational content: it presents the world as being some way, and that presentation can be accurate or inaccurate. Seeing the tomato as red and round is a matter of the experience representing *red* and *round* of the surface in front of you. Illusion is the experience getting the object wrong; hallucination is the experience representing a scene that isn't there at all.

The distinctive claim of the view lies in joining these two. A long tradition treats them as pulling apart — the thought being that if perception *represents*, then what we are immediately aware of is the representation, and the world recedes behind it. Intentionalist direct realism denies the inference — and denies it at the root, rather than by finding content a safer place to stand. What an experience says about the world occupies no position at all: not in front of the tomato, and not off to one side as a medium. The experience presents the tomato by saying how things stand with it, which leaves nothing over to be aware of instead.

The Question It Answers — and the One It Doesn't

It pays to separate two questions that the word "representationalism" tends to run together, because intentionalist direct realism answers only one of them.

The first question is about the *objects* of perception: when I see the tomato, what am I directly aware of — the tomato, or some inner stand-in? Intentionalist direct realism answers: the tomato. Its rivals here are indirect realism and the sense-datum theory on one flank, and naïve realism and disjunctivism on the other.

The second question is about *phenomenal character*: what is the redness of my experience of the tomato — what does that felt quality consist in? The answer the book gives is strong representationalism (equivalently, strong intentionalism): phenomenal character consists in representational content of the right kind. Its rival here is qualia realism, the view that experience carries an intrinsic inner quality over and above any content.

These are different claims with different opponents, and it helps no one to let a single term do both jobs. Someone could accept that we see external objects directly while denying that phenomenal character reduces to content, or the reverse. The book affirms both theses, and they reinforce each other, but they answer to separate arguments and separate objections. When a passage is fighting the sense-datum theorist, it is defending intentionalist direct realism; when it is answering Block's mental-paint objection, it is defending the identity claim about phenomenal character. Keeping the two apart is what keeps either from being charged with an equivocation.

Where It Stands Among the Rivals

Direct realism is a family, not a single view, and its members divide over how to handle the hard cases — illusion, and especially hallucination.

On one flank sits **indirect realism** and its classical engine, the **sense-datum theory**. Faced with the stick that looks bent in water or the vivid hallucination of a pink elephant, the indirect realist concludes that what we are *really* aware of is an inner item that genuinely bears the perceived property — a bent datum, an elephantine image — with the external world known only by inference. This preserves a tidy, uniform account of veridical and non-veridical experience, and it pays for it with a veil: the world drops behind a screen of mental proxies, and skepticism about the external world walks in through the gap.

On the other flank sit the object-involving direct realisms — **naïve realism** and **disjunctivism**. These share intentionalist direct realism's refusal of the veil; they keep the world itself in view. Martin's version makes the perceived object a literal *constituent* of the experience and therefore denies that a matching hallucination shares its fundamental nature. Other relational views allow more commonality. The pressure concentrates on what positive nature the bad case has and how it shares rational force with seeing. Note the consequence for terminology: since disjunctivism is *also* a direct realism, the bare phrase "direct realism" does not by itself pick out the book's view.

Intentionalist direct realism threads between the two. Against the indirect realist, it holds that no inner object is ever an object of awareness: illusion is misrepresentation, not the displacement of awareness from world to mind. Against the disjunctivist, it holds that veridical perception and hallucination *do* share something — a way of representing the world — so hallucination needs no separate metaphysics. Directness survives not because objects are built into experiences, but because nothing in the experience stands between the perceiver and the world.

View	Direct object of awareness	Hallucination
Indirect realism / sense-data	An inner item (datum, image)	Inner object present either way — at the price of directness
Naïve realism / disjunctivism	The external object, as a constituent of experience	Fundamentally distinct on Martin's view; moderate versions allow more commonality
Intentionalist direct realism	The external world, as represented	Inaccurate content — no inner object, no separate metaphysics

Why Content Screens Nothing Off

The standing objection to the view — pressed by Bill Brewer and others — is that representational content, just by being the kind of thing that can occur without its object (as in hallucination), looks like an autonomous mental item, and therefore like a screen. To perceive *via* content, the objection runs, is to perceive only indirectly.

The reply refuses the objection's opening assumption. The objection treats content as an item and asks only where the item stands — and once the question takes that form, no answer helps: an item between perceiver and world screens the world off, and an item that screens nothing off has no work to do. But what a state says about the world never amounts to a thing occupying a position. The meaning of a sentence stands nowhere between a reader and what the sentence concerns; asking whether it sits in the way misconstrues what a meaning is. Content behaves the same way. You are not aware of it in addition to the tomato, nor aware of the tomato by getting past it: the experience presents the tomato by saying how things stand with it, and there the story ends. This is the transparency of experience put to work: attend as hard as you like to your visual experience of the tomato and you keep finding the tomato — the red, the curve, the sheen — rather than an inner surface bearing those qualities in its place. You may also notice the visual mode in which the tomato is presented, but that mode belongs to the complete intentional structure of seeing; it is not a second object or intrinsic mental paint.

The full engagement with the classical arguments for a mental intermediary — illusion, hallucination, perceptual variation, the time-lag of light, and the causal story from neuroscience — belongs to the companion page on direct realism, where each objection is met in turn. The single move behind all five replies is the one named here: content represents the world, and representing the world wrongly, or representing a world that isn't there, never requires a second world of inner objects to be the thing represented rightly.

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 mind's directedness: perceiving, believing, desiring, fearing, and imagining each point beyond themselves to something. Franz Brentano proposed it in 1874 as **the mark of the mental** — the feature every mental phenomenon has and no merely physical phenomenon shares. A rock has no aboutness; a belief about a rock does.

Brentano's 1874 Thesis

In *Psychology from an Empirical Standpoint* (1874), Brentano proposes intentionality as the criterion distinguishing mental from physical phenomena. The key passage: mental phenomena are characterized by "the intentional (or mental) inexistence of an object" — every mental act contains an object *within itself*, not as a separate thing but as an immanent content.

Intentional inexistence does not mean the object fails to exist. It means the object has *inexistence* — existence within the mental act. Brentano inherits this from the Aristotelian-Scholastic tradition (the *esse intentionale* of the medievals): the mind can contain the form of a thing without containing the matter.

Three categories of mental phenomena, each defined by its directedness:

1. **Presentations** (*Vorstellungen*) — the basic layer. Whenever anything is before the mind at all, a presentation underlies it. Seeing red, hearing a note, imagining a centaur.
2. **Judgments** — presentations accompanied by acceptance or rejection. Belief that p, denial that p.
3. **Phenomena of love and hate** — the affective/conative layer: desire, aversion, pleasure, displeasure. (Later work refines this category extensively.)

Every mental act of the second or third kind has a presentation as its foundation. Judgments and emotional states are about something only because a presentation already presents that something.

The Aristotelian Background

Brentano was trained in Aristotle and Aquinas. The doctrine of intentional inexistence descends from the Aristotelian claim that in perception, the perceiver takes on the *form* of the perceived thing without taking on its matter. Seeing red doesn't turn the eye red; the eye takes on the form "red" intentionally, not materially.

Brentano secularizes and psychologizes this: the intentional object exists *within* the act, but not as a real thing within the skull — as a mode of the act's directedness. The act and its intentional content are one complex phenomenon.

Common Misreadings

The Existence Misreading (Chisholm, Quine, Davidson)

Later analytic philosophers read "intentional inexistence" as a claim about the *existence status of intentional objects* — that Brentano committed himself to a second domain of "mental objects" alongside real physical objects. This reading generates puzzles: does the golden mountain exist in the mind? Must we posit Meinongian realms?

Crane (following careful philological work) argues this reading is wrong. Brentano's *in-existence* is about the mode of presentation within the act, not about ontological status. The sentence "I believe in Santa Claus" does not require Santa Claus to exist anywhere — it requires the act of believing to have a certain directedness.

The Linguistic Criteria Misreading

Chisholm tried to operationalize intentionality via linguistic tests (referential opacity, non-existence of objects, and the like). Critics of this program reply that the tests capture downstream symptoms, not the phenomenon itself: Brentano's thesis concerns mental acts directly, not the logic of sentences about them.

Brentano's Change of Mind (1874 → 1911)

By 1911, Brentano had abandoned immanent intentional objects entirely. The late Brentano (the *reist* Brentano) held that only concrete things (*Realia*) exist — no abstract objects, no immanent mental objects. Intentional reference must be understood without positing any object "in" the mind.

This creates a problem: how does a mental act refer to something nonexistent (unicorns, round squares)? The late theory struggles here. But the essential insight survived the revision intact: intentionality is the mark of the mental, and every mental act is directed.

For the physicalist intentionalist, the early Brentano's *diagnostic* contribution (intentionality as the constitutive feature of mind) matters more than his immanent-object ontology, which that tradition does not adopt.

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 against physicalism — the **knowledge argument**, one of the most discussed arguments in philosophy of mind. It stages a scientist who knows every physical fact about color and has never seen one.

The Thought Experiment

Mary is a brilliant neuroscientist confined from birth to a black-and-white room. She learns everything there is to know about the physical facts of color vision — the wavelengths of light, the neural responses of the visual system, the functional roles of every relevant brain state. She knows all the physical facts.

Then she is released. She sees a ripe red tomato for the first time.

Does she learn something new?

Jackson's claim (1982, 1986): yes, she does. She learns what it is like to see red. And since she already knew all the physical facts, what she learns cannot be a physical fact. Therefore physicalism is false.

What Jackson Argued

Jackson's precise formulation (in "What Mary Didn't Know," 1986) against three objections by Churchland:

1. The argument does not rest on limits of imagination. Mary *would* know what it is like, if physicalism were true. She doesn't. Therefore physicalism is false.
2. The argument does not involve a fallacy of intensionality. The question is not what Mary can logically derive, but what she knows. Giving her more logical acumen doesn't fill the gap.
3. The crucial knowledge is about the experiences of *others*, not her own. Before release, she had no experiences of red; after release, she realizes she was missing something about other people's experiences all along.

The Ability Hypothesis (Lewis, Nemirow)

The main physicalist response: Mary doesn't gain *factual* knowledge on release — she gains an *ability*. She learns how to imagine red, how to recognize red, how to remember red. Knowledge-how, not knowledge-that. No new fact, no threat to physicalism.

Jackson's reply: Mary clearly agonizes about whether she has learned something about others' experiences. She is not agonizing about her abilities — those are a known constant. If ability were all she gained, there would be nothing to agonize about.

Why This Matters

Mary's Room serves as the primary pump for the intuition that something "gets left out" of any third-person physical account. Physicalist responses divide into families: deny that Mary learns anything genuinely new, reduce what she learns to ability (Lewis, Nemirow), or grant her new knowledge while denying that it involves new non-physical facts (the phenomenal-concept and representationalist lines).

Jackson himself later changed his mind and moved to physicalism. His mature answer ran through strong representationalism and the new capacities supplied by a new representational state, and he denied that Mary learned a new truth about how the world is. The book's acquaintance-recognition account is more concessive: it grants genuine epistemic novelty without adding a non-physical fact. The author of the field's best-known anti-physicalist argument came to reject its own conclusion — a rare trajectory, and an instructive one.

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 the assumption that empirical knowledge rests on items simply *given* to consciousness — presented directly, infallibly, and prior to all theory. His attack on that assumption, in "Empiricism and the Philosophy of Mind" (1956), reshaped epistemology and philosophy of mind alike.

What the Myth Is

The picture runs so deep in modern epistemology and philosophy of mind that most theories operate inside it without noticing. On it, the given items present themselves to consciousness directly, fully, and prior to interpretation, and they then serve as the foundational evidence base from which all further empirical knowledge is derived by inference.

Different epistemological programs identify different items as the given: sense-data, sensory impressions, qualia, raw feels, immediate phenomenal items, intuitive given perceptual experiences. The specific identification varies; the structural role doesn't. Whatever the given is, it's supposed to do two jobs:

1. Be **immediately apprehended** — non-conceptual, prior to interpretation, available without prior knowledge.
2. Be **epistemically authoritative** — capable of serving as a *reason* for further empirical beliefs.

Sellars's argument: nothing can do both jobs. The Myth is the assumption that something can.

The Structural Diagnosis

The argument runs by examining what each job requires.

To be immediately apprehended, to be there for consciousness without prior conceptual work, an item must arrive before interpretation. It must be pre-conceptual. The peasant looking at an after-image does not need to know any physical theory to have the after-image; the after-image is just *there* for him. This is supposed to be the foundationalist's bedrock.

To serve as a reason for empirical beliefs, to be the kind of thing that can be evidence for a claim, an item must enter into inferential relations with beliefs. It must be the kind of thing that can stand in logical relations to propositions. Reasons are propositional or quasi-propositional; they support claims by entering the logical space of justification.

But nothing can be both pre-conceptual *and* propositional. Anything pre-conceptual is below the level at which justification operates. Anything that can serve as a reason has already been conceptualized — has entered the logical space of reasons, in Sellars's later phrase. The supposed *raw* given is not capable of grounding inference; the supposed *evidence-giving* given is not raw.

The structural argument, then: the Myth depends on an item being simultaneously below the line of conceptualization (to be foundational) and above it (to be epistemically authoritative). No item can occupy both positions, so the foundational role is impossible to fill.

The diagnosis has not gone unchallenged. Foundationalists — Laurence BonJour prominently, after his own conversion from coherentism — reply that direct acquaintance with experience can justify beliefs without itself being conceptual, and the debate over whether anything can play that role remains live.

The Jonesean Myth as Positive Reconstruction

Having demolished the foundationalist picture, Sellars offers a positive story about how thought-talk and sensation-talk actually function. The story runs through a fictional anthropology — the "myth of Jones."

Imagine a community that has only behavioral and observational vocabulary. They describe what people do, what they say out loud, what their bodies are doing. They have no vocabulary for *inner episodes* — no talk of thoughts, feelings, sensations as inner items.

A genius Jones in this community develops a theory. According to the theory, overt verbal utterances are the culmination of a process that begins with *inner episodes* — episodes modeled on the antecedent practice of public discourse. These inner episodes, which Jones calls *thoughts*, are theoretical posits, introduced to explain patterns in overt behavior. They are not discovered by inward inspection; they are inferred from the explanatory work they do.

The community then learns to *report* its own thoughts, but only because the theory has been put in place. Self-reports of inner episodes are not direct readings of pre-given inner items; they are exercises of a theoretical vocabulary applied reflexively.

The Jonesean myth is Sellars's anti-Cartesian alternative to the inner-theater picture. Thought-talk is *theoretical*, modeled on the antecedent practice of public discourse, downstream from speech-talk. The supposed inner items are not directly given; they are conceptual posits.

This positive reconstruction bears directly on the treatment of inner speech (the inner voice as internalized public speech, not as a pre-given inner thing). See *Inner Speech*.

Why the Myth Matters for Philosophy of Mind

Sellars's structural diagnosis is upstream of several moves in contemporary anti-Cartesian philosophy of mind. Three deserve specific naming.

Anti-qualia. Some qualia theories cast phenomenal character as an intrinsic, immediately accessible given, and Sellars's diagnosis applies directly to that foundationalist form. A sophisticated mental-paint theorist can instead treat qualia as intrinsic properties of experiential states without making them perceived inner objects or infallible evidence. Sellars therefore removes one major route to qualia realism; the further identity argument in Chapter 6 must answer the stronger property-based rival. See *Qualia*.

Anti-sense-datum theory. Sense-data are inner colored items, directly apprehended, infallibly characterized — the perceptual version of the Myth. Sellars's structural attack applies straightforwardly: sense-data can't do the epistemic work the theory assigns them. See *Indirect Realism and Sense Data*.

Anti-Cartesian inner theater. The inner theater is the picture in which inner items are presented to an inner observer. The "presented" relation is supposed to be immediate and authoritative — exactly the Mythic role. Sellars's diagnosis is part of the broader case against the inner theater. See *The Inner Theater*.

In each case, the Myth's structure is the same: an item is supposed to be both raw (so it can be foundational) and reason-giving (so it can do epistemic work). Each case fails for the same reason.

What Replaces the Myth

The Myth's collapse leaves a question: what does ground empirical knowledge? Sellars's answer is that empirical knowledge is grounded *holistically* in a network of beliefs, observations, and inferences, none of which has the foundational role the given was supposed to play. Particular perceptual reports are reliable insofar as they are produced by trustworthy mechanisms operating in normal conditions; they are not reliable because they bottom out in self-authenticating inner items.

McDowell develops the line for our time: empirical knowledge involves *both* receptivity (the world impacts us) and spontaneity (our conceptual capacities are involved at every level). There is no level at which receptivity alone, before spontaneity kicks in, delivers a determinate content. See *Phenomenology* for related discussion.

On this picture, phenomenal character is real; introspective reports are reliable; perceptual experience does ground beliefs about the world. None of this requires the Mythic foundation. The patterns of reliability, the holistic structure of justification, the way concepts enter at every level — these do the work the Myth claimed only an inner given could do.

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 at least some fundamental physical entities have conscious experience—that there is something it is like to be one. It need not make rocks, cities, or every composite object conscious. A related view, panprotopsychism, places nonconscious but specially consciousness-constituting properties at the base instead. Once regarded as fringe, both views now receive serious attention as responses to the hard problem. Philip Goff and Galen Strawson defend promi-

nent versions; David Chalmers maps the wider Russellian family and its combination problems without simply endorsing it.

What It Claims, and Why It Tempts

The modern, respectable motivation is **Russellian monism**, a family containing Russellian panpsychist, panprotopsychist, panqualityist, realization, and identity versions. Physics, the thought goes, describes the *structural and dispositional* features of matter—its relations and what it can do—while staying silent on its *intrinsic categorical* nature, or what concretely has those powers. Russellian panpsychism fills that role with phenomenal properties; panprotopsychism uses non-experiential properties specially apt to constitute consciousness. Panqualityism posits qualities without microsubjects, while powerful-quality versions join qualitative nature to causal power. The consciousness-fundamental branches present themselves as a third way between reductive physicalism and dualism; identity and local versions may instead remain forms of physicalism. See *Russellian Monism*, *Dualism*, and *Physicalism*.

The Combination Problem

The standing objection asks how a fundamental phenomenal or qualitative base yields the unified macro-experience of a person. Micro-subjects do not obviously sum into a macro-subject, but that objection reaches only views that posit microsubjects. Panqualityism instead owes quality, awareness, and subject constitution; phenomenal bonding owes a principled constitutive relation; cosmopsychism owes localization and thinning; Tye's consciousness* owes transfer. Chalmers and Goff have made this family of problems the central battleground. See *Russellian Monism* and *The Hard Problem of Consciousness*.

Tye's Conversion

Michael Tye, one of the book's principal representationalist interlocutors, moved to panpsychism in a 2021 book and restated the position in 2024. His route does not run from PANIC alone. He argues that phenomenal consciousness is essentially sharp, introduces a fundamental determinable he calls *consciousness**, and holds that it transfers to a whole once a relevant global-workspace condition is definitely met, while representational content supplies the determinate character. The book disputes the sharpness inference and treats the bearer-transfer as primitive and unexplained. See *Michael Tye*.

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 fundamen-

tal 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 (PCS) is the physicalist's most developed reply to the hard problem of consciousness and the anti-physicalist arguments that feed it. Its claim: we possess special *phenomenal concepts*, ways of thinking about experiences acquired through having them, that stand conceptually isolated from our physical and functional concepts. Two concepts can pick out *one and the same property* while remaining cognitively independent, so the felt unbridgeable distance between "C-fibers firing" and "this — pain" is exactly what we should expect even if they refer to the same thing. On the strategy, the explanatory gap is a fact about our concepts, not about the properties they pick out: real, predicted, and harmless to physicalism.

The Strategy

The strategy is associated with Brian Loar, David Papineau, Katalin Balog, and others, but it names a family rather than one mechanism. Accounts construe phenomenal concepts as recognitional, demonstrative or indexical, quotational or constitutional, and directly referential. Some ground their deployment in undergoing the relevant experience; others specify a different cognitive architecture. The shared physicalist proposal says that phenomenal and physical concepts can remain cognitively isolated while co-referring. No amount of physical description thereby *confers* the first-person route, and zombies may remain conceivable without being possible. See *Mary's Room* and *Zombies and Conceivability*.

How It Answers the Gap

The PCS lets the physicalist *concede the epistemic gap and deny the ontological one*. Mary learns no new fact; she gains a new, phenomenal concept of a fact she already knew physically — a

reading that pairs naturally with Lewis's ability hypothesis and Tye's representational account. The zombie's conceivability shows a gap between two concepts, not two properties. This is the machinery behind the claim that the hard problem is epistemic, not ontological. See *The Hard Problem of Consciousness* and *The Explanatory Gap*.

Chalmers's Objection

The strategy has a formidable critic. David Chalmers ("Phenomenal Concepts and the Explanatory Gap," 2006) requires a topic-neutral psychological feature C to explain a topic-neutral epistemic situation E. If C can vary with the physical facts fixed, the gap has moved into C; if it cannot, Zombie Mary has C too, while her corresponding belief may lack Mary's truth or justification. C then explains shared recognition and gap-talk without explaining their whole epistemic difference. The debate remains live.

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" carries at least three distinct meanings in contemporary philosophy of mind, and they get conflated often enough that any careful treatment has to keep them apart.

Sense 1: phenomenology-as-method — the descriptive enterprise inaugurated by Edmund Husserl and developed by Merleau-Ponty, Heidegger (in a different register), and the broader Continental tradition. It aims to describe the structures of conscious experience from the first-person point of view, prior to and independently of empirical-scientific theorizing. Methodologically: bracket the natural attitude, attend to the structures of givenness, describe what shows itself.

Sense 2: phenomenology-as-data — the lower-case use familiar in analytic philosophy of mind. "The phenomenology of seeing red," "the phenomenology of pain," "the phenomenology of moral disagreement." This sense names *what experience is like* — the explanandum that any theory of consciousness has to account for. No methodological commitment to Husserl follows from using the word in this sense.

Sense 3: cognitive phenomenology — the contemporary debate about whether occurrent conscious thought has its own proprietary phenomenal character, distinct from sensory imagery

and from the phenomenology of attitudes. Pitt, Strawson, Siewert, and Bayne and Montague's volume frame this debate. The question is narrower than either of the first two senses: not whether phenomenology is a legitimate method, not whether experience has phenomenal character at all, but whether *thinking*, as distinct from seeing, hearing, and feeling, carries its own distinctive what-it's-likeness.

Much of the cross-talk between Continental phenomenology, analytic philosophy of mind, and the cognitive-phenomenology literature comes from sliding between these senses without flagging the shift.

The Method and Its Critics

Husserl's tradition insists that any theory of mind must answer to the structures of experience as actually undergone — not as reconstructed from third-person physical descriptions. Brentano, Husserl's teacher, supplies the entry point: intentionality as the mark of the mental. On the tradition's own terms, the directedness of consciousness is a datum, not a hypothesis, and any theory that makes consciousness vanish into computation or behavior has thereby failed.

Naturalists typically part company at the transcendental ambitions. Husserl's epoché, the move from the natural attitude to the transcendental-phenomenological standpoint, is not an innocent methodological step, the naturalist argues: it presupposes that consciousness forms a separate domain investigable independently of its physical realization. On the naturalistic reading, the structures phenomenology describes belong to representational activity in embodied, world-engaged organisms, and they call for empirical explanation rather than transcendental insulation. Husserlians reply that the epoché suspends the natural world without denying it; the dispute remains one of the deepest fault lines between the two traditions.

Merleau-Ponty sits closer to the naturalist. *Phenomenology of Perception* (1945) grounds perception in bodily activity and world-involvement, refuses the Cartesian inner standpoint, and anticipates much of what analytic philosophy now discusses as embodiment.

Phenomenology as Data

The phenomenal character of experience — what it is like to see, hear, feel, think — is the central explanandum for any theory of consciousness. Every party to the contemporary debate accepts phenomenology in this sense; the disputes concern what the data are data *of*. Representationalists such as Tye hold that phenomenal character consists in representational content of a specific kind (poised, abstract, nonconceptual, intentional). Qualia realists read the same data as revealing intrinsic inner properties of experience. Phenomenology-as-data does not by itself referee that dispute; it sets the table for it.

The Cognitive-Phenomenology Debate

Is there something it is like to *think* — over and above the imagery, the inner speech, and the emotional tone that accompany thinking? Defenders (Pitt, Strawson, Siewert) argue yes: the phenomenal character of thought is proprietary, distinctive, and individuative; you know what you are thinking in a way no sensory accompaniment explains. Deniers, including many representationalists, reply that what we report as "the experience of thinking" resolves into sensory

imagery, inner speech, and attitude, with no phenomenal residue left over. The debate turns on introspective claims that have proved hard to adjudicate, which is itself part of the lesson.

Phenomenology and Transparency

A potential tension: phenomenology-as-method seems to require introspective access to the structures of consciousness, while the transparency of experience implies that introspection delivers the world rather than the experience. Does a transparency-based view undermine its own appeal to phenomenology?

The representationalist resolution: transparency does not deny that experience has structures or that introspection can reveal them. It denies that what introspection reveals are inner objects with intrinsic properties. The phenomenologist's question — what is it like to perceive a tree? — is answered by attending to *how the tree shows up*: as having a particular shape, color, size, position, depth. The tree-as-shown-up is the content of the experience, and describing it carefully is describing the experience's phenomenology. What transparency rules out is a different exercise: peeling away the tree to attend to a residue of pure inner appearance behind it. See *Transparency*.

Standing Confusions

The given-as-data confusion. Treating phenomenal data as foundational, self-authenticating, immune from theoretical revision. Wilfrid Sellars pressed against this in *Empiricism and the Philosophy of Mind* (1956): nothing can be both immediate and reason-giving. Phenomenological *description* is not raw data delivery; it is an exercise of concepts — fallible, revisable, and subject to the same standards of inference as any other claim.

The inner-object reading. Treating phenomenology as the description of inner mental items rather than of the world-as-shown-up. Representationalists diagnose qualia realism, the sense-datum picture, and the inner theater as products of this reading; qualia realists deny that it is a misreading at all. See *Qualia*.

The Cartesian misreading of Husserl. Reading transcendental phenomenology as a return to a Cartesian standpoint of pure consciousness, with the physical world bracketed away. This reading makes phenomenology look flatly incompatible with naturalism. The Merleau-Pontian corrective — the embodied, world-involving subject — is closer to what the tradition at its best delivers.

The conflation of senses. Sliding between phenomenology-as-method, phenomenology-as-data, and cognitive phenomenology without flagging the shift. Careful work keeps them apart.

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, in philosophy of mind, holds that the world consists of one kind of thing — causally closed, investigable in principle by the natural sciences — and that conscious experience, mental representation, intentionality, and meaning belong within that one world rather than standing outside it in a second, non-physical realm. On this view there exist no irreducible psycho-physical laws and no non-natural facts about the mental. Whatever the sciences ultimately discover about the structure of nature, mind fits inside that structure. The position dominates contemporary analytic philosophy of mind, uniting thinkers as different as Smart, Armstrong, Lewis, Dennett, Tye, and Dretske despite deep disagreements about what consciousness actually is.

What the Position Claims

A standard way to state the thesis precisely is the supervenience formulation: the physical facts fix all the facts. Any two worlds identical in every physical respect are identical in every respect, mental facts included — no difference in experience without a difference in the physical base. Physicalists divide over whether this supervenience holds with metaphysical necessity (the orthodox view) or merely natural necessity, a weaker claim that shades toward property dualism.

A stronger contemporary formulation adds a clause that does much of the real work in current debates: phenomenal character does not lie at the bottom of things as a further fundamental feature of the world. Many views call themselves "one world" while quietly seating consciousness among the fundamentals; that last clause separates physicalism from them.

What "Physical" Means Here

"Physical" functions as a placeholder for whatever ultimate scientific investigation discovers about the structure of nature. The position does not commit to current physics being the final word; it commits to the methodological premise that whatever the world turns out to be, it is *one kind of thing*, investigable in principle by the natural sciences, governed by laws that do not run out at the boundary of the mental.

This invites Hempel's dilemma: if "physical" means *current* physics, physicalism is probably false, since current physics is incomplete; if it means *some ideal future physics*, the thesis risks emptiness, since no one knows what that physics will contain. Most physicalists accept a version of the second horn while insisting the thesis stays substantive in two ways — it denies a second domain, and in its stronger form it denies that phenomenal character is fundamental. Those denials keep their content regardless of what the final inventory of particles or fields looks like.

Physicalism, Naturalism, and the Reductive Question

"Naturalism" names a close cousin, often used interchangeably, but the terms can come apart. Naturalism commits broadly to one world continuous with the sciences — roomy enough that a property dualist (Chalmers calls his own view "naturalistic dualism") and a panpsychist can both claim the name. "Physicalist" generally serves as the narrower label, pinned to the denial that phenomenal character is a fundamental ingredient of reality.

Physicalism also divides along the reductive axis:

- **Reductive (type-identity) physicalism** holds that each type of mental state is identical to a type of physical state — pain *is* C-fiber firing (Place 1956, Smart 1959, Armstrong). Its classic difficulty is multiple realizability.
- **Non-reductive physicalism** holds that mental properties are physically realized but not identical to, or reducible to, lower-level physical descriptions. Multiple realizability — the same mental state realized in different physical substrates (Putnam, Fodor) — is taken to show that mental kinds cross-classify physical ones, so token physicalism plus realization replaces type-identity.

The reductive/non-reductive split is orthogonal to the question of substrate. Whether mind requires a specific kind of matter (biological neurons) or can be realized in others (silicon) is the substrate-independence question — see *Carbon Chauvinism* and *Functionalism*.

The Standard Anti-Physicalist Arguments

Three arguments target physicalism directly, and the literature carries standard replies to each.

The conceivability argument (zombies). If a being physically identical to a conscious person but wholly lacking experience is conceivable, and conceivability is a guide to possibility, then phenomenal facts do not follow from physical facts, and physicalism is false (Chalmers). Physicalist replies typically deny the inference from conceivability to metaphysical possibility, or locate the apparent gap in our concepts rather than in the world. See *Zombies and Conceivability*.

The knowledge argument (Mary's Room). Mary, who knows every physical fact about color from inside a black-and-white room, seems to learn something new on first seeing red; so some

facts are non-physical (Jackson 1982 — a position Jackson himself later abandoned). Physicalist replies include the ability hypothesis (Mary gains know-how, not a new fact: Lewis, Nemirow) and the old-fact/new-mode reading (one fact grasped under a new, phenomenal mode of presentation). See *Mary's Room*.

The explanatory gap. Even granting a psycho-physical identity, we cannot see *why* a given physical state should feel the way it does (Levine 1983). The dominant physicalist response treats the gap as epistemic rather than ontological — a feature of how phenomenal concepts work, not a hole in nature. See *The Explanatory Gap*.

Chalmers's Taxonomy: Type A through Type F

David Chalmers's map of positions on consciousness ("Consciousness and Its Place in Nature," 2002) has become the standard way to locate any view, and it sorts the physicalist options especially cleanly. It turns on two questions: is there an *epistemic* gap between physical and phenomenal truths, and is there a corresponding *ontological* gap?

- **Type A materialism** denies the epistemic gap. Once the functions are explained, consciousness is explained; there is no further "hard problem." Zombies are inconceivable on reflection, and Mary learns no new fact (analytic functionalism, eliminativism; Dennett, Lewis, the early Harman).
- **Type B materialism** *accepts* the epistemic gap but *denies* the ontological gap. The gap is real — zombies seem conceivable, Mary seems to gain something — yet phenomenal states are still identical to physical states. These identities are necessary but knowable only *a posteriori*, on the model of "water is H_2O." The gap then gets explained not by a gap in the world but by the special character of phenomenal concepts (the phenomenal concept strategy). Type B is the home of *a posteriori* physicalism (Loar, Papineau, Block, Hill, Perry; Tye is often read this way). See *The Phenomenal Concept Strategy*.
- **Type C materialism** holds that the epistemic gap is deep but closable in principle; Chalmers argues it is unstable and collapses into Type A or Type B (Nagel, McGinn, Churchland at points).
- **Type D** is interactionist dualism (two realms in causal commerce); **Type E** is epiphenomenalism (the physical is closed, experience a causally idle extra — Jackson 1982); **Type F** is Russellian monism / panpsychism, which seats consciousness in the intrinsic nature of the physical. See *Dualism*, *Panpsychism*, and *Russellian Monism*.

Chalmers's own view sits at D or F; he presents Type B as the most serious *physicalist* option while arguing it cannot ultimately escape the gap. The taxonomy matters because the difference between Type A and Type B is exactly the difference between *denying* the hard problem and *accepting it as a genuine epistemic problem while denying that it shows anything ontological* — the fault line along which physicalists of consciousness sort themselves.

What Physicalism Is Not

Three recurrent misreadings are worth heading off:

- **Physicalism is not eliminativism.** Most physicalists affirm that experience, intentionality, and meaning are real — physically realized and relationally structured, not explained away. Eliminativism (some mental categories fail to refer) is one species of physicalism, not the genus.

- **Physicalism is not scientism.** The thesis concerns what the *world* is made of, not the claim that current scientific vocabulary captures everything important or that philosophy reduces to science. Conceptual work remains real work.
- **Physicalism is not anti-phenomenology.** The felt quality of experience is a datum any adequate physicalism must respect, not a nuisance to argue away. See *Phenomenology*.

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

Two influential frameworks in contemporary cognitive science claim significant implications for the philosophy of perception. **Predictive processing** says the brain is a prediction engine: rather than passively receiving sensory input, it generates top-down predictions and updates on the error signal. **The interface theory of perception** (Donald Hoffman, University of California, Irvine) says perception is a species-specific "user interface" that represents adaptive fitness payoffs rather than objective reality; evolution selects for reproductive success, not for truth.

Both frameworks sometimes get glossed as showing that perception does *not* put us in direct contact with the world — that we live behind a veil, or inside a "controlled hallucination" (Anil Seth's phrase). Whether that philosophical conclusion follows from the scientific frameworks is a separate question, and it divides even philosophers who fully accept the science.

Predictive Processing

In the predictive-processing framework (Clark, Hohwy, Friston, and the free-energy program), perception is the brain's best top-down hypothesis tested against incoming signals, with prediction error driving revision. As neuroscience the framework has real credentials and may well describe how perceptual systems are organized. Its philosophical interpreters divide sharply. Jakob Hohwy reads it as supporting an inferential, internalist picture: the brain sits behind an evidentiary boundary, inferring a world it never directly touches. Andy Clark reads the same science the opposite way: predicting is how an embodied brain locks onto and engages its world, not a screen between them. Seth's "controlled hallucination" has become the popular banner for the internalist gloss. See *The Inner Theater*.

The Interface (Desktop) Theory

Donald Hoffman's interface theory pushes harder. Evolution, he argues, tunes perception for fitness rather than truth, so our perceived world is a "desktop" of useful icons that systematically *conceals* reality as it is. The icon on your screen is not the file; the tomato in your visual field, on Hoffman's view, stands to reality roughly as the icon stands to the circuitry. The desktop metaphor is the sharpest contemporary statement of the "perception as veil" intuition, which is why it draws direct answers. Critics press two standard replies: fitness-tuned perception must still track the world well enough to act on it, and a global anti-realism about perception threatens to undercut the evolutionary premises the theory itself stands on. See *Indirect Realism and Sense Data.*

The Philosophical Dispute

Does the science force the veil? The case for the veil gloss: if what you perceive is the brain's best guess, or a species-specific interface, then the object of your awareness would seem to be the guess, the model, the icon — not the world.

The case against runs through the vehicle/content distinction. That the brain *predicts*, *models*, and *constructs* describes the **vehicle**: the machinery by which perceiving happens. It does not follow that the model is the **content**, what perceiving is *of*. On this reply, the brain's predicting is what we see *with*, not what we see *at*, and "controlled hallucination" mistakes the one for the other. Perceptual constancies — the achievements by which the system locks onto mind-independent objects across varying input (Burge) — then stand as exactly what a veil picture leaves unexplained. See *Vehicle and Content*, *Transparency*, and *Direct Realism*.

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" is philosophy's name for the intrinsic, subjective, non-relational qualities of conscious experience — the "what it is like" of seeing red, tasting coffee, feeling pain. The redness of red-experience, the painfulness of pain, the specific tang of espresso: on the classic conception, these are properties of experiences themselves, not properties of the world the experiences represent. Whether anything answers to that conception is one of the central disputes in the philosophy of mind.

What Qualia Are Supposed to Be

On the standard qualia-realist picture (Block, Shoemaker, Levine), qualia are:

- **Intrinsic:** they belong to the experience itself, not to any relation between the experience and the world
- **Non-relational:** they do not consist in representing anything; they just *are* what they are
- **Directly accessible:** introspection reveals them immediately and infallibly
- **Ineffable:** they resist complete capture in functional or physical description
- **Private:** only the subject has direct access

The Representationalist Objection

A central line of objection, developed by representationalists such as Tye and Harman, holds that qualia so conceived do not exist. The concept, on this view, involves a systematic reification: a pattern of representational relations gets treated as an inner substance, an entity in its own right.

The move from "experience has phenomenal character" to "experience has intrinsic non-relational inner properties" does not follow, the objection runs. It rests on a level-confusion: the vehicle of representation gets confused with the content represented. The subject describes the world from a first-person perspective, then mistakes that description for a description of an inner entity.

Tye makes the conclusion explicit: if we define qualia as "directly accessible intrinsic qualities of experiences," there are no qualia. What exists instead: qualities of external surfaces and objects, of which the subject is directly aware via experience, and which enter into the representational content of the experience. These may reasonably be called "phenomenal qualities" in a less restrictive sense — but they are not inner things.

The realist does not concede the point. Block's "Mental Paint" (2003) argues that the transparency observations the objection leans on settle much less than they seem to: attending to the world in experience, he contends, remains compatible with intrinsic mental qualities — the paint — partly constituting what the experience is like. What introspection actually reveals is the front line of the dispute. See *Transparency*.

Block's Access/Phenomenal Distinction

Ned Block distinguishes two things "consciousness" can mean:

- **Access consciousness (A-consciousness):** a state is A-conscious if its representational content is available to reasoning, reporting, and the control of action.
- **Phenomenal consciousness (P-consciousness):** a state is P-conscious if there is something it is like to be in it — if it has qualia.

Block argues these can come apart. A state might be A-conscious without being P-conscious (the zombie case), or P-conscious without being A-conscious (certain forms of blindsight or overflow).

Intentionalists need not deny Block's distinction between access and phenomenal consciousness. They deny that phenomenal consciousness must consist in non-representational qualia. Strong representationalism identifies a conscious state's phenomenal character with its complete world-presenting intentional profile — content under a perceptual or bodily-affective mode, with its determinacy, field structure, and attentional organization — not with access as such. Availability, recurrence, and poising remain competing hypotheses about state consciousness, and overflow remains a live empirical challenge to some of those hypotheses rather than proof of intrinsic mental paint.

The Inverted Qualia Scenario

Could two people represent the world identically — respond to red things as red, green things as green, never making errors — yet have phenomenally inverted experiences? What I call "red" feels to you the way what I call "green" feels to me?

Qualia realists answer yes: the scenario is conceivable, and therefore possible.

Intentionalists reply: if the two representational contents are truly identical, including all causal-historical relations to colors, the phenomenal characters cannot differ. On examination, the inverted-qualia scenario either smuggles in a difference at the representational level (different causal histories, different functional roles) or becomes incoherent.

The Absent Qualia Scenario

Could a functional duplicate of me — a system realizing the same functional organization — lack qualia entirely? Block uses this case to argue that qualia are not exhausted by functional organization.

The intentionalist reply: a genuine functional duplicate, with the same world-involving causal history and environmental relations, would have the same representational content and thereby the same phenomenal character. "Functional duplicate" in the relevant sense cannot be specified without including the causal-environmental grounding — which is precisely what the absent-qualia scenario abstracts away.

The Sellarsian Backdrop

Sellars's attack on the Myth of the Given (*Empiricism and the Philosophy of Mind*, 1956) supplies the structural backdrop for the contemporary anti-qualia line. The Myth is the view that empirical knowledge rests on epistemically self-authenticating immediate experiences — items simply *given* to consciousness, fully present, infallibly known. Qualia, on the standard realist picture, look like exactly this kind of item: intrinsic, immediately accessible, infallibly introspected.

Sellars's structural diagnosis: nothing can be both immediate (non-conceptual, available without prior knowledge) and capable of serving as a reason, hence inferentially related to beliefs. Whatever consciousness delivers, it does not deliver self-authenticating inner objects ready to ground further claims about them. When qualia realists insist that phenomenal character must be intrinsic, non-relational, and directly accessible, critics charge them with reinstating the Myth of the Given in the philosophy-of-mind register. The proposed fix runs the same in both registers: reject the picture of self-authenticating inner items, and relocate the work to relational, world-directed properties of the conscious state.

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 is the position that qualia exist — that conscious experiences have intrinsic, non-relational, introspectively accessible qualitative properties, and that these properties are genuine features of reality rather than confusions, illusions, or mere ways of talking. Where the concept of qualia names the candidate items, qualia realism is the metaphysical thesis that the items are real. It ranks among the most widely held positions in contemporary philosophy of mind, and it is the chief rival to representationalist accounts of phenomenal character.

What the Realist Claims

The qualia realist holds that when you see red, taste coffee, or feel pain, your experience instantiates a property — the redness of the red-experience, the painfulness of the pain — that is:

- **Intrinsic:** it belongs to the experience itself, not to any relation the experience bears to the world;
- **Non-representational:** it is not exhausted by, or reducible to, what the experience represents;
- **Introspectively accessible:** the subject is directly and immediately aware of it;
- **Real:** it is a genuine feature of the world that any complete account of the mind must accommodate, not eliminate.

The realist need not be a dualist. The position comes in physicalist as well as anti-physicalist forms. What unites them is the insistence that phenomenal character is a *something*, a property in its own right, that cannot simply be paraphrased away into functional or representational facts.

The Main Varieties

Physicalist qualia realism. Qualia are real and intrinsic but identical to, or realized by, physical properties of the brain. Ned Block's notion of "mental paint" (the intrinsic phenomenal properties that color our mental states) is the most influential development. Block holds that mental paint is real and not representational, while remaining officially neutral on whether it is physical. Sydney Shoemaker likewise treats phenomenal (or "qualitative") properties as real properties grasped through introspection. His account fixes them by their functional and causal profile rather than eliminating them by it.

Dualist or "phenomenal" realism. David Chalmers uses the term *phenomenal realism* for the thesis that conscious experiences have phenomenal properties that are real and not reducible to the physical. On this version, the reality of qualia, combined with the explanatory gap, is taken to push toward property dualism: phenomenal properties are further fundamental features of the world, over and above the physical ones. This is the realism the hard problem is built on.

Anti-physicalist arguments for realism. Frank Jackson's knowledge argument ("Mary's Room") and the conceivability of zombies are standardly read as arguments *for* the reality of qualia and *against* their being captured by physical or functional description. Jackson himself later abandoned the dualist conclusion; the arguments remain central to the realist case.

The Arguments That Motivate It

Qualia realism is rarely defended by a single positive theory — as Averill and Gottlieb note, it is "rarely (if ever) fully developed in positive terms." Its force comes mainly from a cluster of

arguments meant to show that *something* must be left over once all the functional and representational facts are fixed:

- **The knowledge argument:** Mary learns something new on first seeing red, so there is a fact about the phenomenal that the complete physical story omitted.
- **The explanatory gap (Levine):** even a complete physical account leaves it unexplained *why* a given physical state feels the way it does — suggesting an unaccounted-for qualitative fact.
- **Inverted and absent qualia (Block):** two systems could share functional organization while differing in, or wholly lacking, qualia — so qualia are not exhausted by function.
- **Conceivability of zombies:** a physical duplicate of you with no experience at all seems conceivable, hence (the argument runs) phenomenal properties are something further.

The Realism / Anti-Realism Landscape

Qualia realism sits between two opposing rejections. **Eliminativism** (Dennett, and on some readings Rey) denies that qualia exist at all: the concept is too confused, or too laden with incoherent commitments (intrinsicness, infallibility, ineffability), to pick anything out. **Representationalism** or **intentionalism** (Harman, Tye, Dretske, Byrne) takes a middle path. It preserves *phenomenal character* — there really is something it is like — while denying that this character consists in intrinsic non-relational properties. On this view, phenomenal character is the *representational content* of experience, so the realist is right that experience has genuine qualitative character and wrong about where that character lives. The dispute over transparency is the front line: realists read introspection as revealing intrinsic mental qualities; representationalists read it as revealing only the world as represented.

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 explana-

tory 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 Naturalism.

Reduction and Elimination

A physicalist can say two very different things about the mind, and almost everything turns on which one it comes to. The first says that experience, intentionality, and meaning are perfectly real, and then tells you what they *consist in* — the physical, relational, embodied facts that make them what they are. The second says that some piece of our mental vocabulary answers to nothing in the world at all, and should be dropped the way we dropped *phlogiston* and *caloric*. The first reduces; the second eliminates. Both keep to one world, physically closed, so from a distance they look like the same view. They are not, and a great deal of confusion about physicalism comes from running them together. Reducing a thing preserves it. Eliminating it denies it was ever there.

What Reduction Preserves

To reduce a phenomenon is to say what it *is* in more basic terms while granting, throughout, that it exists. Water reduces to H_2O; the reduction did not abolish water or reveal that your thirst was an illusion. It told us what water is. Lightning reduces to atmospheric electrical discharge; heat in a gas consists in the mean kinetic energy of its molecules. In every case the everyday thing survives the reduction intact — indeed the reduction is *about* it, and would be false if the thing did not exist. What the reduction removes is not the phenomenon but a certain picture of it: that it must be a further, free-standing ingredient over and above the physical facts that realize it.

Applied to the mind, this pattern gives the identity claim its shape. The felt character of seeing red is real; it consists in your experience representing the world in a certain embodied, world-directed way. The claim does not say the redness is a fiction. It says the redness is not a private inner object glowing behind the seeing — it is that world-presenting activity, described from the inside. The reductive physicalist therefore stands committed to the reality of what she reduces: the whole point of reducing rather than eliminating.

What Elimination Denies

Eliminative materialism makes the opposite move. It holds that a stretch of our mental talk is the vocabulary of a *false theory* — a folk theory that will not survive a mature science of the brain, and whose terms, like *phlogiston*, turn out to name nothing. Paul Churchland argued this for the propositional attitudes: beliefs and desires, he suggested, belong to a primitive "folk psychology" that a completed neuroscience might simply replace, at which point there would be, strictly, no such things as beliefs. On the phenomenal side, some deflationists press the analogous line about

qualia — that the very notion has grown so tangled with bad philosophy that the right conclusion says there are no qualia, not that qualia are physical.

The eliminativist and the reductionist agree on the metaphysics — one world, physically closed — and disagree on the bookkeeping. Where the reductionist says "this is real, and here is what it consists in," the eliminativist says "this was never real; here is the better vocabulary that replaces it." Elimination is not a stronger form of physicalism; it is a *different verdict* about which of our concepts refer.

Why the Distinction Is Easy to Lose

The trouble: a reductive identity statement, read carelessly, sounds exactly like a debunking. "Phenomenal character is just representational content" — the little word *just* does the damage, and so does the bare copula. Say "pain **is** C-fiber firing" and a listener hears: so pain is *nothing but* wiring, so the hurt was never really there. The sentence invites the deflationary reading even when no deflation is meant. This is why the careful move reaches past the flat verb *to be* for a verb that names the actual relation — *consists in*, *is realized by*, *is constituted by*. "Pain consists in a certain damage-tracking representational state realized in the nervous system" cannot be misheard as "there is no pain." It says what pain is, and in saying so affirms that there is pain to say something about. Choosing that verb does more than decorate; it keeps a reduction from being mistaken for an elimination.

A Note on the Word "Reduction"

One further tangle trips up even careful readers. "Reduction" gets used in two senses. In one sense it means *a priori* conceptual analysis — the claim that mental truths follow, by meaning alone, from a full physical description, so that a sufficiently clever reasoner could deduce the one from the other with no residue. In the other sense it means *a posteriori* ontological identity, on the model of water and H_2O: the identity holds, and holds necessarily, but is discovered rather than deduced, and no armchair analysis delivers it. These two come apart, and much confusion follows from not noticing that they do. One can deny the first while affirming the second. Reductive representationalists like Tye show the combination in action: the representational identity is fully reductive, and yet no *a priori* deduction carries you from the physical facts to the feeling, so a real epistemic gap stays open. Denying *a priori* reducibility is not eliminativism, and it is not property dualism either. It is what accepting the gap as epistemic, and only epistemic, actually looks like.

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 is the relation between a word (or a thought) and what it is about. It looks like the simplest thing in semantics — the word *cat* picks out cats — and the simplicity is a trap. The *referential theory of meaning* takes the relation to be the whole of meaning: a word means by standing for its referent, like a label on a thing. Two classical results break this picture. What emerged from the wreckage, in the externalist tradition, is a view of reference as a world-embedded achievement rather than a tag: a relation established through causal history and community practice, not readable off the symbol itself.

The Referential Theory and Its Collapse

Mill gave the cleanest version: a proper name is a bare mark whose whole meaning is the object it denotes. Two classic results break it. **Frege's puzzle**: "Hesperus is Phosphorus" is informative while "Hesperus is Hesperus" is trivial, though both names pick out Venus — so meaning outruns reference; something more (Frege's *sense*, the mode of presentation) rides with the word. **Russell's empty names**: "the present King of France" and "Pegasus" are perfectly meaningful with nothing to refer to — so meaning cannot *be* the referent.

Sense and Reference

Frege's repair distinguishes *Sinn* (sense, the way an object is presented) from *Bedeutung* (reference, the object itself). This is the indispensable correction to the naive theory. The externalist tradition presses past an internalist reading of that repair: sense and reference are not labels stored inside a symbol and readable from its form. Expressions without standalone referents can contribute

to meaningful acts within practices whose use reaches a world, while names reach their bearers through causal history and community participation. The error lies not in making meaning world-involving, but in treating the world-involving relation as a tag legible off the token.

Descriptivism vs. Direct Reference

Two accounts compete over how a name reaches its bearer. **Descriptivism** (Frege–Russell): a name abbreviates a description the speaker associates with it. **Direct reference** (Kripke, *Naming and Necessity*): a name rigidly designates its bearer with no descriptive content, fixed by a causal-historical chain running back to an initial baptism and sustained by speakers' referential intentions. Kripke can *look* like a revival of the naive referential theory, but he relocates reference *off* the symbol and into a world-involving practice — confirming, not refuting, the externalist point.

Reference as a World-Involving Achievement

Once you ask *how* reference gets fixed, the answer comes out externalist: causal-historical relations to the world, not anything inside the speaker (see *Semantic Externalism*). A causal chain alone may still leave a particular address underdetermined. Concrete *de re* reference requires a discriminating world-involving relation somewhere in the practice's ecology or genealogy: a producer anchors the particular, and consumers can inherit that handle through testimony, deference, and public use. Original mechanism-level content requires a correctness standard independent of an interpreting practice, not a proprietor's act of appropriation. Whether the larger system itself refers, believes, or understands remains a further question about integrated consumer use. See *Intentionality* and *Teleosemantics*.

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 externalism holds that the content of mental states (what they are *about*, what they refer to, what would make them true) depends partly on factors outside the thinker's skull. Two people can be intrinsically identical — same brain states, same internal physical organization — while differing in what their thoughts are about, because they stand in different relations to their environments. The slogan, from Hilary Putnam, is direct: "meanings just ain't in the head."

The position emerged from two landmark arguments: Putnam's Twin Earth thought experiment (1975) and Tyler Burge's arthritis case (1979). Together they established that both natural-kind terms and ordinary mental concepts are individuated partly by environmental and social factors that lie outside the thinker. The view has since ramified through philosophy of mind, philosophy of language, and cognitive science.

The Wittgensteinian Precursor

Putnam and Burge did not invent the externalist intuition; they sharpened a thought already present in Wittgenstein's private-language argument (*Philosophical Investigations* secs. 243–315). The argument holds that meaning cannot be fixed by what goes on inside a single skull, because following a rule (applying a term consistently to the same kind of case) requires the possibility of public correction. Without that, there is no fact of the matter about whether the next application agrees with the prior ones.

Wittgenstein's target is narrower than Putnam's — it concerns the constitution of meaning rather than its individuation across communities — but the structural point holds: a privately-anchored content is not a content at all. Wittgenstein gives the externalist line its foundational motivation; Twin Earth and the arthritis case give it its modern analytic apparatus. The two work in tandem. The private-language argument makes the case that meaning *cannot* sit purely inside the head; Twin Earth and the arthritis case show *what kind of external relations* fix it instead.

Putnam's Twin Earth

The thought experiment (Putnam, "The Meaning of 'Meaning'", 1975):

Twin Earth is a planet exactly like Earth in every observable respect, with one exception. The liquid that fills Twin Earth's lakes, falls as rain, comes out of taps, and is called "water" is not H_2O. It is a different chemical compound, call it XYZ, that shares all the macroscopic properties of water (colorless, tasteless, odorless, drinkable) but differs in its microstructure.

Now consider an Earthling and their Twin-Earth counterpart in 1750 — before chemistry advanced enough for either to know the microstructure of the liquid they call "water." They are, we may stipulate, molecularly identical. Same brain states, same internal organization, same functional role for the concept "water." And yet:

- The Earthling's word "water" refers to H_2O
- The Twin-Earthling's word "water" refers to XYZ

The conclusion: what a term refers to (its extension) is not determined by what is "in the head." The two speakers have the same internal states but different semantic contents. Meaning depends partly on the actual causal-historical relation between the speaker and the stuff in their environment.

Putnam's slogan: "Cut the pie any way you like, 'meanings' just ain't in the head."

Natural kind terms — water, gold, tiger — get their extensions fixed by their microstructural essence and the causal-historical chain connecting current uses to original samples, not by any internal description the speaker associates with them.

Burge's Social Externalism

The arthritis case (Burge, "Individualism and the Mental", 1979):

Suppose a man, call him Arthur, has a variety of arthritic complaints. He also has a pain in his thigh and mistakenly believes he has arthritis in his thigh. (Arthritis, by definition, affects joints; the thigh is not a joint.) Arthur has a false belief involving the concept ARTHRITIS.

Now consider a counterfactual situation: Arthur has the same physical history, the same internal states, the same dispositions — but lives in a community where "arthritis" covers any rheumatic ailment, including those affecting the thigh. In this community, Arthur's thigh complaint would count as arthritis. The counterfactual Arthur's belief "I have arthritis in my thigh" would be *true*.

The conclusion: the content of Arthur's belief (which concept ARTHRITIS picks out) depends on the linguistic conventions of his community. Two individuals with identical internal states can differ in what concepts they exercise. This extends Putnam's naturalistic externalism, grounded in physical microstructure, to a social externalism grounded in linguistic community — internal organization underdetermines content either way.

Tye on Memory Externalism

Tye ("Externalism and Memory", 1998) extends Twin-Earth externalism to memory contents. If I am slowly and unknowingly transported to Twin Earth after living on Earth, my word "water"

eventually comes to refer to XYZ rather than H_2O as I assimilate into the Twin-Earth linguistic community. At that point, my *memories* of drinking "water" in my youth (memories that originally had Earth-water as their content) now have Twin-water, "twater," as their content — even though my internal states haven't changed.

The upshot: memory contents, like belief contents, are individuated partly by environmental and socio-environmental factors. What I remember answers to the environment in which the remembering takes place, not purely to what is stored inside my skull.

Tye also argues that memory *images* are transparent in exactly the way perceptual experiences are: introspecting a memory image of red apples discloses the apples as red, not some inner image with intrinsic redness. The photographic model of memory images fails both introspectively and empirically. Memory, like perception, is world-directed, not inner-object-directed.

Narrow Content and the Internalist Response

Some philosophers (Fodor, early Loar) respond by distinguishing:

- **Wide content**: the externally individuated content that picks out H_2O vs. XYZ — different for Earthling and Twin-Earthling
- **Narrow content**: whatever is shared by the two — perhaps a function from contexts to wide contents, or a Fregean sense abstracted from reference

The externalist reply: narrow content is either too thin to do psychological work (it can't explain why the Earthling reaches for water and not twater) or it collapses back into functional role without grounding. The interesting explanatory notion of mental content, the one that connects minds to the world, is wide content. On the externalist view, narrow content is a philosopher's abstraction, not a psychologically real kind.

Privileged Access and Externalism

If content is partly environmental, how can we have privileged access to our own thoughts? I can know what I'm thinking without investigating my environment — yet what I'm thinking depends on my environment.

Tye's response (1998): externalism is compatible with privileged access because privileged access is access to *what one is thinking now*, and what one is thinking now is externally individuated. I have direct access to the fact that I am thinking *about water* (H_2O); my twin has direct access to the fact that they are thinking *about twater* (XYZ). Each has privileged access to their own thought; the thoughts differ because the environments differ.

The apparent paradox dissolves: privileged access doesn't require that content supervene on internal states. It requires only that thinkers have direct, non-inferential access to the contents they have — whatever those contents turn out to be.

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

The Distinction
A *simulation* is an implementation used as a model of something. A *realization* is an implementation that possesses the property at issue. The two can come apart, but the label *simulation* does not tell us that they do.

A weather model does not move air or rain on the desk. A digestion model does not break down protein. Those properties require physical powers the computer lacks. A simulated calculator, however, really calculates, and a simulated controller wired to a boiler really controls it. The

diagnostic question is therefore not *is this called a simulation?* It is *which powers does this physical implementation actually realize?*

The application to mind is conditional. If consciousness depends only on an intrinsic causal-computational organization, a system realizing that organization is not disqualified because it also models a brain. If consciousness requires additional world-involving and norm-fixing powers, those must be realized too. The hurricane analogy shows that structural fidelity does not transfer every property automatically. It does not identify the missing property for mind.

Intrinsic Implementation

Putnam- and Searle-style triviality arguments defeat unconstrained computational mappings: with enough interpretive freedom, a wall can be described as running any finite program. They do not defeat disciplined implementation.

Chalmers requires counterfactual-supporting causal structure. Ma and Kanai's intrinsic computational functionalism strengthens the demand: the organization must be specifiable without an observer's labels, invariant under structure-preserving relabeling, and displayed in the system's intervention profile. A system meeting those conditions genuinely realizes the causal-computational organization. This is an observer-independent physical fact, not a story laid over the machinery.

The remaining dispute concerns semantics and consciousness, not whether the implementation is real. Intrinsic transition structure alone does not fix which worldly conditions its states are about. A critic of simulated consciousness must identify the further consciousness-relevant organization rather than repeat the word *simulation*.

The Candidate Stake

Mind, Matter, and Meaning defends thermodynamic self-maintenance coupled to adaptive regulation as a candidate basis of **autonomy, practical interest, and welfare**. The system's activity must help preserve the physical organization doing the activity. Its regulatory states must distinguish movement toward and away from viable conditions and vary the response. That organization institutes a descriptive functional standard: some trajectories support the continuation of the regulatory organization and others undermine it.

The claim does not derive a categorical ought from causal success. Plenty of things persist without regulating their own persistence; a candle manages the first and fails the second. In adaptive reciprocal closure, regulatory activity counterfactually affects viability, while departures from viable conditions alter regulation. That two-way dependence makes the success condition internal to the organization being maintained. "Norm" here names the resulting possibility of function and malfunction, not a moral command written into matter.

Selection or training can give a submechanism content. Whether that content counts as original depends on whether the correctness standard arises independently of an interpreting practice. The stake does not create, elevate, or own that content. It changes what success and failure mean for the continuing organization: correct tracking may now guide activity on which its persistence

depends, and failure may harm it. Content explains what a state says; stake helps explain why getting it right matters to this organized life.

This yields five constitutive questions. Selection, learning, and causal history can produce **mechanism-level content**. Flexible use guided by that content can support **domain-specific competence**. Several capacities constraining one another in an answerable system can support **partial understanding**; breadth, persistence, and cross-domain traffic make that understanding more integrated and world-involving. A supported threshold relation marks **state consciousness**, while integrated persistence identifies the subject and the world-presenting profile determines character. Reciprocal self-maintenance can separately establish **autonomy, interests, and welfare**. Reasons-responsive agency cuts across the map when represented considerations govern conduct through the persisting control organization. None of the later achievements mints content, and self-maintenance is not a prerequisite of consciousness.

In organisms, metabolism realizes thermodynamic self-maintenance. Metabolism is the biological instance, not the substrate-general requirement. Life and carbon are not necessary. An artificial or virtual system could acquire a comparable stake if its physical implementation closed the same regulatory loop. It could acquire content through selection, learning, or world-involving history whether or not it crossed that further line.

The System and Its Platform

The system boundary follows reciprocal maintenance at a stated time scale, not the casing, corporate ownership, or the location of a process. The time scale must match the processes that constitute, regulate, and repair the organization; it cannot be chosen merely to secure a preferred verdict. A component belongs to the candidate system when its activity helps maintain the relevant closure and its continued operation depends on that closure. A platform component that technicians, schedulers, or an independent power system keep available regardless of the agent's success remains outside the loop.

This criterion permits nested and distributed systems. A robot may include remote components; a factory, colony, or firm may qualify as a self-maintaining organization with practical interests at that level. Such cases do not automatically produce a unified subject that understands or feels.

Why Current Systems Lack This Stake

The text-trained and agentic architectures the book analyzes commonly represent goals, energy budgets, health variables, and even their own continued operation. They may also contain genuinely content-bearing submechanisms. Their platforms nevertheless maintain the physical implementation independently of the agent's success. The scheduler, data center, power system, and technicians keep the process available while the agent tends represented set-points. The viability drama and the continuation of the realizing machinery come apart. That supports a bounded verdict about autonomy and welfare, not a semantic or consciousness verdict.

That diagnosis stays bounded to those architectures; it does not survey every public or possible AI system. A digital system is physical all the way down. Couple its adaptive activity to energy acquisition, repair or reconstitution of its realizing machinery, maintenance of a boundary, and world-tracking that bears on those activities, and the case changes.

Copyability and Backup

Copyability is not a criterion of content or consciousness. A backup may reveal that a current agent leaves its continuation to infrastructure outside its loop, but the architecture does the work, not the file.

- A non-copyable process may fail to maintain itself; the candle dies quite unbacked-up.
- A self-maintaining system would not lose its norm because an engineer learned to copy it.
- A duplicate raises questions about identity, succession, and inherited history. It neither creates nor removes the original's stake.

Geoffrey Hinton's *mortal computation* remains important because it proves that computation can bind tightly to the detailed physics of one device. Hardware-specificity is neither necessary nor sufficient for a stake. It is an engineering clue, not the semantic criterion.

Virtual Worlds

Virtual objects and events can be real digital objects and events. Their causal relations can matter to an agent. Designer assignment does not permanently condemn every category inside a made world to borrowed meaning; designers also build laboratories in which organisms acquire content of their own.

A bare web of formal relations, without selected producer-consumer use, leaves distal reference unfixed. Training within a virtual environment can give mechanisms content when their consumers use those states and differential retention preserves the arrangement. Whether that content counts as original depends on whether the correctness standard remains borrowed from an interpreting practice. Causal-historical traffic can further fix worldly reference. Neither achievement alone establishes understanding by a unified agent.

A practical stake requires more: the content-bearing activity must contribute to maintaining the organization that sustains and revises it. A represented health variable supplies no such relation by its name. A physical implementation whose self-maintenance depends on tracking virtual conditions could acquire autonomy and welfare, even though designers built the world and its physics. Its content would still come from the relevant selection, learning, and causal history.

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

How do the symbols in a symbol system get their meaning, if their definitions are only more symbols? Stevan Harnad named the problem in 1990. Look up a word in a monolingual dictionary and you get more words; look those up and you get more words still. A purely symbolic system is a closed loop of tokens defined by other tokens — a "merry-go-round" that never touches the world. Unless some symbols are connected to what they are about by something other than further symbols, the whole system is ungrounded: formally manipulable, but meaningless. The problem is the precise, contemporary form of the worry the Chinese Room dramatizes, and it bears directly on large language models.

Harnad's Statement

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 that connects at least some elementary symbols to the categories of things the system can pick out in its perceptual interaction with the world ("sensorimotor toil"), so that higher symbols built from them ("symbolic theft") inherit a connection to the world. A symbol is grounded, in his minimal sense, if the system can pick out what it refers to. See *The Chinese Room* and *Computation and the Church-Turing Thesis*.

Grounding Is Not Yet Meaning

Harnad himself is careful here: he says he has offered "a theory of symbol grounding, not a theory of meaning." Grounding gets a symbol connected to a category in the world; whether the grounded system thereby *means* anything, or has any inner life, is a further question grounding alone does not settle. Both halves hold together: grounding is *necessary* for meaning (an ungrounded symbol system means nothing on its own), but not obviously *sufficient*. See *Embodiment*.

The LLM Case

For a bare text-only model considered apart from persistent memory, tools, and world-coupled scaffolding, Harnad's problem remains sharp. Distributional form alone does not fix which distal worldly thing a token concerns. But the model's word-types descend from public linguistic practice, and training can make inherited content causally internal and explanatorily indispensable

inside its mechanisms. Corpus-mediated producer-consumer chains may also transmit public reference. Mallory's word2vec analysis concerns a more limited target: actual word-types and the contexts in which they occur. It treats those linguistic facts as features of the world that trained mechanisms can represent. Multimodal contact, tools, action, and world-anchored feedback can add new reference-fixing relations, though whether a deployed product or integrated agent uses them as an understander remains a separate question. See *Semantic Externalism* and *Vehicle and Content*.

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 selected function and producer-consumer organization. What a state represents depends partly on what states of that kind were stabilized to do, not simply on whatever happens to cause the state now.

The theory's central achievement is to make error intelligible. Smoke correlates with fire and tree rings with age, but correlation alone does not say what a signal is supposed to get right. A producer-consumer history can privilege one correspondence rule over its rivals and explain how a state can be false even while remaining causally effective.

The Frog

A frog's visual circuitry responds to small dark moving things. Does its state represent *flies*, *edible moving prey*, or *small dark moving objects*? Current cause cannot decide. Teleosemantics asks which mapping mattered to the stabilization and successful use of the producer-consumer organization.

The answer may remain coarse. If *prey at this location* and *fly at this location* explain the organization equally well, the grain may be indeterminate while both readings exclude a BB pellet. If *nutritive moving prey* and *small dark moving thing* remain tied, the extension is unsettled and so is the verdict that the BB was misrepresented. Relevant interventions help: hold darkness fixed while varying nutritiveness, then reverse the test. A mapping earns content when such differences predict consumer success and explain the organization independently of an observer's preferred vocabulary.

The content belongs first to the producer-consumer mechanism. That does not turn the mechanism into a tiny frog or settle whether the whole animal perceives a distal object. Burge's objectification threshold and teleosemantic content answer different questions.

Function, Content, and Level

An **etiological function** depends on history: doing F helps explain why a trait or state type was selected, learned, or differentially retained. A correspondence rule maps a family of producer states onto conditions that consumers use. The history makes that rule explanatory rather than optional.

Three questions must remain separate:

1. Does the producer-consumer mapping fix a correctness condition?
2. At what level should that content be attributed?
3. What larger capacities does the system's use of the content justify attributing?

Selection and learning may answer the first question at a subpersonal or mechanism level when the success condition itself fixes what the states concern. They can also merely internalize a semantic standard supplied by an interpreting practice. Wider coordination across memory, inference, learning, conflict resolution, and action may justify system-level belief or understanding. It does not create the component's content.

Ownership Withdrawn as a Semantic Condition

The book previously added a proprietor requirement: an owner had to appropriate a historically shaped state before its content could become original. That requirement is withdrawn. It answered a question about whose state or capacity we should attribute to the larger system, not what makes the state accurate or inaccurate.

Self-maintenance remains important to theories of autonomy and agency. It may identify an individual with practical interests and explain persistence or welfare. It does not supply a missing semantic relation. Add a self-repair loop while holding the producer-consumer mapping fixed and the represented condition need not change.

Major Positions

Millikan. Proper functions belong to historically reproduced producer-consumer systems. Consumer use helps determine satisfaction conditions and makes misrepresentation possible.

Dretske. Learning can recruit an informational indicator for a use. The acquired function lets representational content diverge from present correlation.

Papineau and Neander. Their accounts develop the role of successful action, selected function, and natural information in fixing nonconceptual content.

Hutto and Myin. Radical enactivists deny that covariance and function yield truth-conditional content. The reply begins with a nonsemantic job description: one distal-condition-sensitive variable remains portable across consumers, recombines with changing inputs, and gets corrected at the mapping even when final payoff is rescued elsewhere. Ward's decouplability test helps distinguish that organization from a one-link controller. Hutto and Myin can still call the richer

system teleosemiotic rather than semantic. The book takes the history-privileged correspondence to constitute a satisfaction condition because it unifies portability, recombination, correction, and error under one rule. The disagreement concerns that semantic verdict, not a hidden proprietor.

Mallory. Training supplies a selection history for word2vec's producer-consumer mechanisms. On his account, an embedding represents a fact about a word-type and its contexts while directing the downstream mechanism to produce a distribution over context words. Actual linguistic distributions, rather than the numerical code alone, supply the target against which training corrects the output. He argues that these states carry original content and explicitly treats linguistic relations as part of the world. This claim concerns mechanisms and their linguistic targets, not whole-model understanding or reference to the objects named by the words.

Swampman

Swampman at the instant of creation lacks the history teleosemantics uses to fix content. His self-maintaining organization cannot substitute for that missing history. A molecule-for-molecule duplicate can therefore lack teleosemantic content at $t = 0$ even though he immediately has the machinery needed to begin acquiring it.

A trained language model presents a different profile: it has a training history, but a text-prediction history may internalize semantic norms inherited from language rather than originate them. It may also lack perception, distal reference, integrated belief, or understanding. The contrast shows why causal history, semantic independence, autonomy, and system-level mentality must not be collapsed into one test.

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

In 1985 the Oxford physicist David Deutsch, a founder of quantum computation and a constitutional optimist about what physics permits, noticed something the standard Church-Turing thesis had quietly smuggled in. The ordinary thesis identifies *effective calculability* with what a Turing

machine computes: a claim about functions and procedures, mathematical through and through. Deutsch's move was to ask what *physical* fact would have to hold for that mathematical claim to bear on the actual world. His answer he called the **Church-Turing principle**: "every finitely realizable physical system can be perfectly simulated by a universal model computing machine operating by finite means."[1]

Read that twice, because the reframing does real work. The original thesis talks about abstract computability; Deutsch's principle talks about *physical systems* and their *perfect simulation*. It asserts that the world contains a universal simulator — a machine that, fed the right program, can reproduce the behavior of any finitely realizable chunk of physical reality. That converts a theorem of recursion theory into a hypothesis about the furniture of the universe. Deutsch then argued that classical physics, with its continuous dynamics, does *not* underwrite the principle, whereas quantum theory, via the universal quantum computer, does. The principle and the physics rise or fall together.[2]

Why does any of this matter for philosophy of mind? Because the principle, if true, looks like exactly the bridge the computationalist has always needed. The brain is a finitely realizable physical system. By the principle, it can be perfectly simulated by a universal computing machine. And a perfect simulation of a thinking thing, the computationalist concludes, would itself think. The Church-Turing-Deutsch (CTD) principle promises to ground in physics what mere recursion theory could never deliver: a *physical* license for the leap from brain to program to mind.

The Computationalist's Fallback

The standard case for computationalism leans on the bare Church-Turing thesis, and that lean has a well-known weakness. As Piccinini argues at length, "the brain is computable" and "the brain is a computer whose mind is its software" are simply different claims; the first can hold while the second fails, and treating the second as a corollary of the first is the *Church-Turing fallacy*.[3] A computationalist who has read Piccinini knows the mathematical thesis won't carry the weight alone. CTD is where she retreats to higher ground.

The fallback runs like this. The earlier argument failed because it tried to get an empirical conclusion about minds out of a thesis about abstract functions. CTD repairs the gap by being *itself* an empirical, physical principle: it does not merely say a function is computable, it says a physical system is *perfectly simulable* by a universal machine. So the chain no longer limps across a category boundary: physical brain → physical principle of universal simulation → faithful simulation of the brain → mind. Every link, the computationalist now insists, lives on the physical side of the ledger — exactly the upgrade the Church-Turing fallacy seemed to demand.

And CTD is not a fudge. It is a serious, physically motivated thesis from one of the architects of quantum computing, and it really does close the *first* gap, the one between mathematical computability and physical simulability. The computationalist who reaches for it has correctly diagnosed where the cheaper argument broke. The open question is whether the repair reaches far enough.

The Simulation-Realization Reply

Critics answer that it stops one seam short. Granting CTD in full — granting that the brain can be perfectly simulated by a universal quantum computer — buys the computationalist a conclusion about *simulation* and nothing more. And simulation is precisely the layer at which critics of computationalism have already dug in.

The middle plank of the case against machine understanding is the distinction between simulation and realization. A perfect computational description of a process is not yet an instance of that process. A simulated hurricane produces no rain and wets nothing; a simulated digestion breaks down no food. Those examples mark the difference between describing causal powers and possessing them. But the word *simulation* carries no verdict by itself. A silicon or virtual system may realize the relevant organization rather than merely depict it. CTD guarantees simulability; it does not decide whether a particular implementation realizes the powers relevant to mind.[4]

So the CTD fallback, the critic concludes, overshoots its target. It establishes simulability and then helps itself, without argument, to the further claim that the simulation *is* a mind. But "perfectly simulable" and "such that its simulation thinks" are as different as "perfectly simulable" and "such that its simulation is wet." The hard question — whether running the structure on alien hardware constitutes the phenomenal, world-directed mental life of the thing simulated — is exactly the question CTD does not touch. The computationalist has climbed to higher ground only to find the same chasm in front of her.

There is also a more direct objection aimed at CTD's own footings, from the philosopher of physics Christopher Timpson. Timpson argues against Deutsch that the Church-Turing hypothesis is *not* in fact underwritten by a physical principle — that Deutsch has run together two genuinely distinct claims, the (mathematical) hypothesis about computable functions and the (physical) "Turing Principle" about universal simulation, and that the latter does not do the foundational work Deutsch assigns it.[5] If Timpson is right, the fallback collapses earlier still: CTD does not even secure the physical grounding it advertises, let alone the leap from grounding to mind.

The Quantum Dimension

The quantum angle is the part of Deutsch's paper that draws the most attention, and it is tempting to think it changes the philosophical stakes. On the critics' analysis it does not, and seeing why is clarifying.

Deutsch's universal quantum computer extends what can be *efficiently* simulated, and it is the reason classical physics fails the principle while quantum theory satisfies it. One might hope that this quantum upgrade smuggles in something a classical Turing machine lacks — irreducible physical concreteness, perhaps, or some special intimacy with the brain's actual substrate — and that this something is what realization required. But the hope misreads what the quantum computer adds. It extends the range and efficiency of simulation; it does not by itself establish realization. A universal quantum computer modeling a hurricane is no more meteorologically wet than a classical one running the model more slowly, but a concrete quantum system could realize some target organization if it possessed the relevant causal powers. The implementation has to earn that verdict independently.

The quantum dimension therefore does not settle the dispute. It widens the class of systems a universal machine can faithfully model while leaving open whether a particular implementation realizes the content-bearing, integrative, and phenomenal organization required for mind.

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 strong computationalism. 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.

Notes

1. 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 p. 99. Deutsch explicitly contrasts this *physical* formulation with the standard mathematical Church-Turing thesis, which he regards as a special, less committal case.

2. Deutsch, "Quantum Theory," 100–101. Classical physics fails the principle because its dynamics are continuous while the simulating machine operates by finite means; the universal quantum computer Q is introduced precisely as the device under which quantum theory satisfies the

principle in its strong form. Note that Q extends the *efficiency and range* of simulation without breaching the Turing barrier for recursive functions — it computes no non-recursive function.

3. Gualtiero Piccinini, "Computationalism, the Church-Turing Thesis, and the Church-Turing Fallacy," *Synthese* 154, no. 1 (2007): 97–120. Piccinini distinguishes the empirical hypothesis that the brain performs computations from the much stronger claim that the mind is the software the brain runs, and shows that arguments deriving the latter from the Church-Turing thesis are unsound.

4. The simulation/realization distinction is developed at length in *Simulation and Realization*; the canonical examples (a simulated hurricane is not wet; a simulated digestion digests nothing) trace to Searle's insistence that a formal model of a causal process is not an instance of that process.

5. Christopher G. Timpson, "Quantum Computation and the Church-Turing Hypothesis," ch. 6 of his DPhil thesis (Oxford, 2004); published in revised form as "Quantum Computers: The Church-Turing Hypothesis Versus the Turing Principle," in *Alan Turing: Life and Legacy of a Great Thinker*, ed. Christof Teuscher (Berlin: Springer, 2004). Timpson, a philosopher of physics, argues that Deutsch conflates the (mathematical) Church-Turing hypothesis with a distinct (physical) "Turing Principle," and that the former is not underwritten by the latter.

The Explanatory Gap

The explanatory gap is Joseph Levine's name for a puzzle that survives even a true physicalism: granting that mental states are brain states, we still cannot see *why* a given brain state should feel the way it does. Levine introduced the term in "Materialism and Qualia: The Explanatory Gap" (1983). The move is subtler than Chalmers's later hard problem and worth keeping distinct from it.

Levine's Specific Move

Levine grants the physicalist that mental states might *be* brain states. The identity claim — pain = C-fiber firing, or whatever the right neural correlate turns out to be — might be true. The question Levine asks is not metaphysical but explanatory: even granting the identity, we cannot *see why* a brain state with that physical description should *feel like that.* The identity, if it holds, holds as a brute fact. We cannot derive the phenomenal character from the physical description.

This is the explanatory gap: a gap between the identity claim (metaphysical) and our ability to explain why the identity holds (epistemic). The physicalist might be right that only one kind of thing exists — the physical — and an explanatory gap would still stand between the third-person physical description and the first-person phenomenal character it identifies with.

Levine's point is therefore *not* that physicalism is false. He explicitly stops short of that conclusion. His point: physicalism, even if true, leaves something unexplained — why the physical realizer feels the way it feels rather than some other way, or no way at all.

Distinguishing the Gap from the Hard Problem

Chalmers's hard problem (1995) is often presented as continuous with Levine's explanatory gap, but the two moves differ in ambition.

	Levine (1983)	Chalmers (1995)
Type of gap	Epistemic	Argued toward ontological
Conclusion	Physicalism unexplained	Physicalism inadequate; property dualism
Argumentative use	Diagnostic of a residual mystery	Premise in a metaphysical argument
Stance toward physicalism	Withholds judgment	Rejects

Levine treats the gap as a feature of our conceptual access; Chalmers treats the gap as evidence that phenomenal properties cannot be reduced to physical properties — and infers non-physical (or at least non-reductively-physical) phenomenal facts. A type-B physicalism accepts something close to Levine's framing while rejecting Chalmers's escalation. See *The Hard Problem of Consciousness* for the full Chalmers engagement.

The reason to keep the two distinct: Levine's argument is much harder to dismiss. Granting the gap as an epistemic feature, while denying that it shows what Chalmers claims, is the physicalist's middle position. Confusing the two moves makes that physicalism look as though it dismisses Levine, when in fact it accepts Levine and rejects only Chalmers's inference from Levine.

The Phenomenal Concept Strategy

One major physicalist reply to Levine is the *phenomenal concept strategy*. Versions appear in Loar, Papineau, Balog, and others; Tye's mature materialism retains acquaintance while rejecting the need for a proprietary class of phenomenal concepts.

The label covers several accounts: recognitional, demonstrative, quotational or constitutional, and direct-reference theories. Their common physicalist use is to explain how first-person and physical concepts can remain cognitively isolated while referring to one state. The book adopts a narrower mechanism: acquaintance with an occurring world-presenting state grounds recognitional capacities that description alone does not confer.

By undergoing an experience, a subject enters an acquaintance relation with occurring world-directed content and gains a recognitional capacity: a way of re-identifying the presented property across later occasions and, potentially, in other perceivers. No third-person description by itself confers that capacity, though descriptions can prepare and refine it. Acquaintance fixes a present reference; it does not reveal an infallible inner sample.

For knowers with our architecture, descriptive and acquaintance-grounded routes can therefore remain conceptually isolated while reaching one physical fact. Mary's release supplies the canonical illustration. She already knows every physical fact about red; seeing red gives her an acquaintance-grounded way of representing and later recognizing a property she previously knew only through description. She gains a real cognitive capacity, not a new non-physical fact. See *Mary's Room*.

Where the Strategy Stops Short

The phenomenal concept strategy answers Levine in his own terms but leaves residual work. Two pieces stand out.

Why these concepts have this structure. A complete account would explain why phenomenal concepts are the way they are — why they require acquaintance, why they don't reduce to functional descriptions. On an intentionalist development of the strategy, the answer comes from intentionalism plus embodiment: phenomenal concepts are recognitional capacities for representational contents that are themselves world-directed and physically realized. The concepts inherit their structure from the contents they represent. This runs deeper than the bare phenomenal-concept reply.

Why the gap feels metaphysical. Even after the epistemic reply is accepted, the *feeling* that more than an epistemic gap remains tends to persist. The intentionalist diagnosis traces this feeling to an artifact of the inner-theater picture: so long as phenomenal character gets implicitly conceived as an intrinsic inner property, the gap between description and acquaintance feels like a gap in nature. Drop the picture, recast phenomenal character as representational content, and the residual feeling has less to feed on. The gap becomes a fact about *how the content is accessed*, not about *what kind of content it is*. See *The Inner Theater* and *Intentionalism*.

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 of consciousness is David Chalmers's name for the question that seems to survive every scientific explanation of the mind: why is there *something it is like* to see, hear, and feel? Explaining how the brain discriminates stimuli, integrates information, and controls behavior looks tractable. Explaining why any of that processing comes with experience looks like

a different kind of question altogether. Chalmers introduced the terminology in "Facing Up to the Problem of Consciousness" (1995), and it has organized the debate ever since.

Chalmers's Distinction

Easy problems: Explaining cognitive functions — attention, memory, reportability, the integration of information, the control of behavior. These count as "easy" not because they are simple but because they are tractable: we know what kind of explanation we want (a functional-mechanistic account) and we have the methods to pursue it.

The hard problem: Explaining why there is *something it is like* to undergo these processes. Why does physical processing give rise to subjective experience at all? Why doesn't it all happen "in the dark"? Even a complete functional account of vision leaves open why seeing red feels like anything.

Levine's Narrower Claim

Levine (1983) made the related point about the **explanatory gap**, and his point is subtler than Chalmers's. He does not claim physicalism is false — only that an explanatory gap stands. Even if we knew that pain = C-fiber firing, we would still face the question of why C-fiber firing should feel like *that*. The identity might be necessary, but we cannot see why it should be. The gap, for Levine, is epistemic, not ontological.

The Type-B response accepts the epistemic gap and denies that it marks an ontological one. The surrounding phenomenal-concept literature offers several mechanisms. The book's narrower account appeals to acquaintance-grounded recognition: description does not confer the first-person capacity acquired by undergoing the relevant world-presenting state. That difference in access leaves one physical fact, not two properties.

The Intentionalist Objection

Intentionalist physicalists argue that Chalmers equivocates on what "explanation" requires. When he says that functional accounts leave something out, he is right — *if* phenomenal character is conceived as an intrinsic inner property. But that conception is precisely what is at issue. On this objection, the hard problem assumes what it needs to argue.

More specifically: Chalmers's conceivability argument for property dualism (philosophical zombies are conceivable, therefore phenomenal properties are not physical) requires that we can coherently conceive of a creature physically and functionally identical to us but lacking phenomenal experience. The intentionalist reply grants the conceiving and denies the inference. A duplicate sharing not just the physics but the world-involving causal history that shaped it carries the same representational content, and thereby the same phenomenal character — so the scenario, however vividly imagined, describes no possibility. The qualifier does real work: on this account it is history, not organization alone, that fixes content, which is why a *functional* duplicate stripped of that history is a genuinely different case.

What the Deflationary View Does Not Claim

The deflationary view does not dismiss the intuition that drives the hard problem. It takes seriously that phenomenal consciousness presents a distinctive explanatory challenge — more so

than, say, explaining digestion. The move is to explain *why* it seems so difficult, not to wave away the seeming.

The argument: the seeming difficulty arises from treating phenomenal character as an inner property rather than a representational relation. Once that correction is made, everything that seemed obvious in lived experience is kept — the richness, the immediacy, the first-person character — while the metaphysical picture that converts those features into evidence for property dualism is rejected.

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 the picture according to which the mind is a private inner space, inside which mental items appear — sensations, images, ideas, qualia — and these items are presented to an inner observer who watches them. On this picture, conscious experience is a kind of internal show: the world causes events out there, those events produce representations in here, and the conscious subject is the spectator of the inner display.

The picture is so deeply embedded in ordinary thinking about the mind that most people never notice they are using it. It surfaces every time someone says "what I see is in my head," "the image of the apple in my mind," "the inner sensation of pain," or "the screen of consciousness." The vocabulary makes the picture sound like a discovery rather than a metaphor.

A long anti-Cartesian tradition treats the inner theater as the central misleading picture that philosophy of mind inherits from Descartes and never quite throws off. Almost every problem in the field (qualia, the hard problem, indirect realism, the explanatory gap, the homunculus problem) gets diagnosed, on this tradition, at least in part as an artifact of this picture. One caution up front: few philosophers defend "the inner theater" under that name. The label belongs to the picture's

critics, and defenders of qualia or sense-data typically deny that their views amount to it — part of the debate is whether the description fits.

The Cartesian Origin

Descartes did not invent the inner theater as a concept, but his framework made it nearly inevitable. The Cartesian division of mind and matter into two substances, *res cogitans* and *res extensa*, required some way to explain how the mind has knowledge of the body's world. The answer that emerged across Descartes, Locke, and the British Empiricists: the mind doesn't perceive the world directly; it perceives *ideas* that the world causes in it, and from those ideas it infers the world.

The picture has four structural features that recur in every later version:

1. **Mind as container.** The mind is a space inside which mental objects reside.
2. **Experience as object.** Experiences are things we encounter, not relations we stand in.
3. **Perception as two-stage.** First we encounter the inner object, then we infer the outer world.
4. **Privileged access as infallibility.** We cannot be wrong about inner objects, because inner objects are constituted by how they appear.

Each feature is sufficient on its own to generate philosophical problems that wouldn't otherwise exist: the problem of the external world, the problem of other minds, the problem of phenomenal properties, the problem of error. The picture creates them and then offers itself as the only framework in which to solve them.

Why the Picture Persists

The inner theater is not just an inherited mistake. It survives because something about ordinary phenomenology seems to invite it. When you close your eyes and imagine a red apple, *something* is going on, and it feels natural to call that something an inner image presented to your awareness. When you have a sharp pain, the pain feels like a thing — a particular, locatable, qualitatively rich item — rather than like a relation or a process. The picture takes these legitimate phenomenological observations and converts them into ontological commitments.

The work for any anti-Cartesian view is to preserve what is right in the phenomenology while rejecting the ontological reading. There is something it is like to imagine a red apple; there is something it is like to feel pain. The picture's mistake is not noticing those facts. The mistake, on the critics' diagnosis, is treating *something it is like* as *something inside the head, presented to an inner observer*. That conversion is the move that has to be undone.

Dissolving the Picture

The intentionalist attack comes at the inner theater from multiple directions, each of which removes one of its structural features.

Transparency presses against feature 2 (experience as object). Introspection ordinarily returns awareness to the world presented rather than revealing a second inner bearer of its colors and shapes. Reflective attention may also disclose the intentional mode, determinacy, attention, blur, or other features of *how* the world is presented. None supplies the private exhibit the theater

requires. With no inner object for a spectator to watch, the picture loses its central image. See *Transparency*.

Intentionalism removes feature 3 (perception as two-stage). Phenomenal character consists in *how* a system represents the world, not in inner properties of the representing state. The two stages collapse into one: representational engagement with the world *is* the experience, not the result of an experience-of-an-inner-item plus an inference. See *Intentionalism*.

The vehicle/content distinction removes feature 1 (mind as container). The neural vehicle of representation is in the head; the content is about the world. Conflating the two — thinking that because the vehicle is in the head the content must be too — is the level-confusion that makes the container picture seem inevitable. Brain states realize perception of the apple; they are not what we are aware of when we perceive the apple. See *Vehicle and Content*.

Sellars's attack on the Myth of the Given removes feature 4 (privileged access as infallibility). Nothing can be both immediate (non-conceptual, available without prior knowledge) and reason-giving (the kind of thing that can serve as evidence). The supposed infallible inner items can't do the epistemic work the picture assigns them. See *The Myth of the Given*.

The four moves are independent and convergent. Each one alone makes the inner theater harder to defend; together, the critics argue, they make it untenable.

The Inner Theater in Contemporary Disguise

The picture has not disappeared from modern philosophy of mind; it has acquired better vocabulary. Three contemporary disguises worth naming:

Qualia realism. When qualia are conceived as intrinsic, non-relational, directly accessible inner properties of experiences, critics read qualia realism as the inner-theater picture with the inner items renamed: the subject as inner observer, the qualia as what the observer apprehends. The intentionalist attack on qualia realism is partly a rejection of this dressed-up Cartesianism — a characterization qualia realists themselves resist. See *Qualia*.

Neuroscience's brain-as-screen. When neuroscience-flavored writing describes perception as the brain *constructing* an internal model that consciousness *has access to*, the inner theater has been recast in updated vocabulary. The neural model plays the inner object; consciousness plays the spectator. Dressed in empirical clothes, the picture inherits the same problems it inherited in the eighteenth century — and this is the version the transparency literature diagnoses most often.

The Cartesian Theater that Dennett attacks. Dennett's *Consciousness Explained* names the picture explicitly and demolishes one version of it. The intentionalist line sides with Dennett's diagnosis of the picture but parts from his eliminativist conclusion — Dennett throws out phenomenal character along with the theater. The alternative move keeps phenomenal character and recasts it as representational content. The theater goes; the experience stays.

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 is one of the oldest and most contested concepts in philosophy. When we ask what the self *is*, the tempting answer pictures an inner owner — a subject behind the experiences, the one who has them, the thing introspection should be able to catch. Whether any such inner item exists, and what it would consist in if it did, has been debated from Descartes through Hume, Kant, and William James into the contemporary philosophy of personal identity.

The Ego Theory and the Bundle Theory

Two broad traditions compete over the question — a fork sharpened by Tye (*Consciousness and Persons*, 2003), following a long line back to Hume and Kant. The **Ego Theory** holds that a persisting subject (a self, an ego) underlies and unifies experience as a real metaphysical entity, distinct from the bundle of perceptions. The **Bundle Theory**, associated with Hume, denies the Ego: there is only the bundle of perceptions, memories, and dispositions, organized but ownerless. Hume gave the bundle its founding statement. Entering most intimately into what he called himself, he caught only particular perceptions, never the self that was supposed to have them.

Each tradition carries its own difficulties. The Ego Theory must explain the nature of the ego and its relation to body and brain; the Bundle Theory must explain what unifies the bundle without appeal to an owner. See *Anti-Reification* and *David Hume*.

The Binder Problem

The bundle theory's standing difficulty is what unifies perceptions into one mind and one continuing life. The answer is not another item inside the theater. The book's provisional criterion identifies the minimal subject as the **relatively maximal persisting control organization** within which contents directly update a shared economy of attention, expectation, memory, goals, correction, planning, and action. Boundaries are tested by intervention: partition the candidate and ask where corrections stop propagating, memories cease to guide the same decisions, plans no longer constrain later action, and conflicts no longer press toward resolution.

The Assembled Self

The book distinguishes a minimal subject from the narrating personal self. The minimal subject is the relatively maximal persisting control organization within which contents can directly update a shared economy of attention, expectation, memory, goals, correction, planning, and action. Boundaries are tested by intervention rather than skin, chassis, or ownership. The narrating self is a further, socially assembled achievement built from language, memory, interpretation, and avowal. Dennett's "center of narrative gravity" is its nearest contemporary relative, while Galen Strawson's Episodic counterexample shows that not every life must be experienced as one continuous story. Practical stake adds vulnerability and welfare; it does not create the subject. See *Daniel Dennett*.

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 Alan Turing's proposed replacement for the question "Can machines think?": a conversation game in which a machine counts as thinking if human judges cannot reliably tell it from a person by typed exchange alone. This page sets out the test, what Turing did and did not claim for it, the standard objections, and where the arrival of large language models leaves it.

The Test

Alan Turing opened "Computing Machinery and Intelligence" (1950) by refusing the question he had been handed. "Can machines think?" he judged "too meaningless to deserve discussion" — the word *think* carried too much disputed freight to settle anything. So he proposed a substitute he could actually run: the **imitation game**.

A human judge sits at a terminal and exchanges typed messages with two hidden partners — one human, one machine. Each tries to convince the judge that *it* is the human. The judge's job is to tell them apart. Turing's wager was that a machine which reliably defeats the judge (one that cannot be distinguished from a person by conversation alone) should be credited with thinking. We need not peer inside it; we need only watch what it can do.

The move is deliberately and self-consciously **behavioral**. It trades an unanswerable question about inner states for an answerable question about performance. And it is restricted, by design, to language: the terminal hides the body, the voice, the face, so that nothing but the quality of the conversation can move the judge. Turing predicted that by the year 2000 machines would play the game well enough that "an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning."

What Turing Was — and Was Not — Claiming

The test is routinely misremembered as a theory of consciousness. It is nothing of the kind, and Turing said so. Replying to "the argument from consciousness" — the objection that a machine could never *feel* anything — he declined to settle whether the machine has any inner life at all. He bracketed the question. His test asks whether a system *thinks* in the operational sense of conversing indistinguishably from a person; it does not ask, and does not pretend to answer, whether there is anything it is like to be that system.[1] He offered an operational criterion for a deflated notion of thinking; his successors inflated a passed test into proof of an inner life — exactly the inference Turing himself withheld.

Turing also anticipated, and answered in turn, a battery of objections he saw coming: the theological objection (thinking requires a soul), the mathematical objection (Gödel's results limit what any formal machine can prove), and "Lady Lovelace's objection" — that a machine "can only do what we know how to order it to perform," and so can never originate anything. His replies were sharp enough that the paper still reads as the founding document of the field.

The Standard Objections

Searle's Chinese Room (1980). The most famous reply is built as a deliberate dual of the imitation game. Searle imagines himself locked in a room, manipulating Chinese symbols by a rulebook well enough to pass a Turing Test for Chinese — and understanding not a word. If a system can pass the test while understanding nothing, then passing the test is not sufficient for understanding. Behavioral indistinguishability and genuine semantics come apart. (See *The Chinese Room.*)

Block's psychologism (1981). Ned Block presses a different lever. Imagine a machine that passes the test by sheer brute lookup — a vast table pairing every possible conversational input, up to some length, with a sensible canned reply. Such a machine could in principle win the imitation game, yet it computes nothing, understands nothing, and possesses no intelligence in any sense we care about; it merely indexes a finite list someone else wrote. Block's point is that the Turing Test is *behaviorist* — it fixes on the input-output profile and is therefore blind to *how* that profile is produced. But intelligence is partly a matter of internal organization, not just external performance. A test that cannot tell a thinker from a jukebox tests the wrong thing.[2]

The general charge. Both objections converge on one diagnosis. The imitation game is a criterion of *seeming*, and seeming is cheap. It can be manufactured by understanding, or faked by lookup, or stumbled into by statistics — and the test, watching only the output, cannot distinguish the cases. Whatever we most wanted to know when we asked "Can machines think?" is precisely what an operational test of conversation is structured to leave out.

The Large-Language-Model Moment

For half a century the test sat as a thought experiment. It no longer does. Contemporary language models often produce text that human judges struggle to distinguish from human writing; on important bounded versions, the game is substantially won. This has not settled the philosophical question — it has sharpened it.

A base model trained on text predicts tokens from a corpus descended from human linguistic practice. Training can make derived content causally internal, and deployed systems may add memory, tools, sensors, and world-sensitive feedback. Bender and Koller's point remains: linguistic form alone does not establish independently grounded meaning or distal reference.[3] Passing the imitation game therefore provides evidence of conversational competence while leaving the provenance and scope of content, understanding, consciousness, and the bearer unsettled. The test succeeded as an engineering target and failed as a self-sufficient criterion of mind.

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.

Notes

1. The bracketing is explicit in Turing's reply to "the argument from consciousness." He concedes he is not concerned to establish that the machine *feels*, only that it can do the relevant work; the question of inner experience he treats as separable from the question of thinking. Strawson presses exactly this point: those who read the imitation game as a test for consciousness mistake Turing, who "explicitly puts aside the question of consciousness" (*The Consciousness Myth*, 2015). The misreading is the rule, not the exception, which is part of why the test does so much rhetorical damage downstream.

2. Block's "Psychologism and Behaviorism" (1981) distinguishes intelligence as behavioral capacity from intelligence as the right kind of information-processing. His "Aunt Bertha machine" — a finite but astronomically large table of canned conversational replies — would, for any bounded conversation, pass the test while doing nothing that deserves the name *thinking*. The lesson is that the Turing Test cannot, in principle, separate intelligence from its imitation, because it fixes on the output and is silent about the process. This is the objection's structural kinship with Searle's: both insert a case where behavior and understanding pull apart, and both indict any criterion that watches only behavior.

3. Bender and Koller (2020) argue that exposure to linguistic form alone does not establish independently grounded meaning or communicative intent. Their octopus thought experiment shows why conversational fluency cannot by itself demonstrate world-grounded understanding.

The present book allows a text-trained mechanism to internalize derived content inherited from linguistic practice and to display domain-specific competence guided by that content; what form-only training does not establish is original content, integrated, world-involving understanding, or consciousness.

Transparency

Transparency names a stable first-person datum: when we introspect perceptual experience, attention ordinarily returns to the presented world rather than revealing a second inner bearer of its colors and shapes. Attend carefully to seeing a red apple and you find the apple, its redness, shape, and surface. Reflective attention may also disclose *how* it is presented — its perceptual mode, determinacy, field structure, attention, blur, or apparent doubling. What does not appear is a private inner exhibit or an independently identifiable coat of mental paint.

The Transparency Thesis

When you try to turn attention inward to catch the "feel" of seeing red, you end up attending outward to redness as a quality of the surface before you. G.E. Moore noticed this first: "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" (1903). Gilbert Harman made it philosophically central in 1990. Michael Tye developed it into the primary argument for strong representationalism.

Tye's Eight-Step Argument

Tye lays out the transparency argument systematically (in "Representationalism and the Transparency of Experience"):

1. In visual experience, the qualities of which you are directly aware appear as qualities of external surfaces — not as qualities of your experience.
2. To suppose otherwise would convict experience of massive systematic error. That is not credible.
3. Introspection of experience yields no inner object — only the external scene and its apparent qualities.
4. This holds even in hallucination: even then, no inner surface is presented, only apparent external qualities.
5. Phenomenal character necessarily changes when the directly-aware qualities change.
6. Therefore phenomenal character consists in representational content of a certain sort — content into which external qualities enter.
7. There are no qualia in the traditional sense (directly accessible inner qualities of experiences). There are only qualities experienced as belonging to external things, which enter into representational content.
8. This generalizes across all modalities: vision, hearing, smell, touch, bodily sensation, emotion.

Key Objections and Responses

Blurry vision objection (Boghossian & Velleman, Block): Seeing something blurrily seems phenomenally different from seeing something as blurry, without any representational difference. Tye's response: seeing blurrily involves *less determinate* representational content (the experience

fails to specify exact contours), while seeing something *as blurry* involves accurate representation of indistinct edges. Different representational contents, no inner residue.

Inverted qualia: If two people represent the world identically but their experiences are phenomenally inverted, transparency fails. Strong representationalism replies that inverted qualia scenarios, on examination, smuggle in the very inner-theater picture they are supposed to demonstrate.

Disjunctivism: Some argue perception and hallucination have nothing in common, so transparency cannot generalize from one to the other. Tye holds that the common-sense and scientific view — that the same conscious state type can occur in both — is far better motivated than disjunctivism.

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 Distinction

A representation has two aspects that are easy to conflate. The *vehicle* is the physical thing doing the representing — a piece of paper, a stretch of ink, a pattern of neural activity, a configuration of pixels. The *content* is what the vehicle represents — the Eiffel Tower, the proposition that snow is white, the red apple on the table. The vehicle and the content are not the same kind of thing and live in different places.

A photograph of the Eiffel Tower is a piece of glossy paper coated with photochemicals (vehicle) that depicts a Parisian iron lattice (content). When you look at the photograph, you attend to what it depicts: the Tower, the Seine, the sky. You do not attend to the paper's chemical composition, the angle of its surface, or the wavelength of light reflecting from its emulsion. The vehicle is transparent to its content. You see *through* it to what it represents.

Perception works through a physical vehicle whose content presents a world — red apple, wooden table, open window. Vehicle-level facts alone do not determine phenomenal character. Within a state already counted as conscious and borne by an identified subject, the book identifies phenomenal character with the complete world-presenting profile: independently fixed content under a perceptual or bodily-affective mode, with determinacy, field structure, and attentional organization. Content fixing, state consciousness, and identification of the persisting bearer remain separate questions. Introspection ordinarily finds the apple's redness, roundness, and glossiness, and may disclose the mode of their presentation; it does not present the neural vehicle as a second perceptual object.

This distinction is one of the most reusable diagnostic tools in the philosophy of perception. It cuts through a wide class of confused arguments about perception, consciousness, and the relation between mind and brain.

What Goes Wrong When the Distinction Is Missed

Three classes of error come from confusing vehicle with content.

The container fallacy. Because the vehicle is in the head, the inference goes, the content must also be in the head. So perception is awareness of inner items, not of external objects. This is the engine of indirect realism. The fallacy is that *vehicle location does not constrain content reference*: a photograph sitting on your desk depicts Paris without Paris being on your desk. A neural state in your visual cortex represents the apple without the apple being in your visual cortex. See *Indirect Realism and Sense Data*.

The neuroscience version of the container fallacy. "What we perceive is a brain state, because the proximate cause of experience is a brain state." Causal proximity is not the same as perceptual aboutness. Dretske's analogy: to say we read ink because ink is the proximate cause of reading is to confuse the vehicle of the written words with what the words are about. The causal chain *enables* perception without replacing its objects. The vehicle is neural; the content is the world.

The inner-object reading of phenomenology. When you have a visual experience of red, *something* is going on, and it feels natural to call that something an inner colored item. But the experience's character is determined by its content, not by intrinsic properties of the vehicle. The phenomenology returns the apple's redness as a property of the apple, not as a property of the experience or of a neural state. Treating the felt redness as a property of the vehicle is the move that reifies qualia. See *Qualia*.

Each of these errors is the same level-confusion in a different costume. The vehicle/content distinction names what is being confused.

Vehicle Location vs. Content Reference

The crispest way to hold the distinction is this:

	Vehicle	Content
Location	Inside the head	About the world
Type	Physical, neural, particular	Representational, semantic, relational
Caused by	Sensory transduction + neural processing	What the vehicle represents
Accessible to	Neuroscience	Phenomenology, semantics

The two columns describe different aspects of the same representational activity. They are not in competition. The vehicle is real and physical; the content is real and world-directed. Trouble starts only when one column tries to do the other column's work.

Why the Distinction Belongs to Intentionalism Especially

Intentionalism — the thesis that phenomenal character consists in representational content — needs the vehicle/content distinction to make the thesis sound. Without it, the intentionalist looks vulnerable to the immediate objection: "you say experience consists in representational content, but representational content can exist when there's no object — as in hallucination — so content is just an inner item, and you've reinstated the inner theater."

The vehicle/content distinction blocks this objection — though not by relocating the content somewhere harmless. Notice what the objection takes for granted: that content amounts to some item, leaving only the question of where the item sits. Content answers to no such description. What a representation says never amounts to a further thing standing anywhere; the meaning of a sentence occupies no position between a reader and what the sentence concerns. Experience works the same way. You experience the apple, not the content, and the experience's presenting the apple to you consists in its having that content — no further step remains for content to mediate.

In hallucination, the content fails to correspond to any external object — there is no apple — but the content is still about an apple in the relevant sense, not about the neural state that realizes the experience. Misrepresentation requires no inner item, only a content whose accuracy conditions don't match the world.

This is what lets intentionalism preserve direct realism in the veridical case (the content reaches the world) while still accommodating hallucination as a genuine phenomenal state (the content runs even without its object). See *Direct Realism*.

The Screened-Off Hand Objection

The sharpest pressure on the vehicle/content distinction comes from screened-off-hand cases. Imagine a hand behind a screen, hidden from view. A brain stimulation produces in you a visual experience as of a hand exactly where the actual hand is. The experience accurately matches the world. Yet the experience is not *of* the hand in the sense direct realism wants — the hand is not the appropriate cause of the experience.

The intentionalist reply uses the vehicle/content distinction to specify *the right kind* of causal-representational relation. Genuine perception requires that the vehicle bear the right causal-historical connection to what the content represents. Deviant causal chains produce experiences with accurate content but without the kind of relation that constitutes perception. The vehicle/content distinction lets us say what the screened-off hand case lacks: not accuracy of content, but the right kind of vehicle-to-world coupling.

This is how the intentionalist handles the screened-off hand objection: the distinction specifies the vehicle-to-world coupling the deviant case lacks.

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 being physically and functionally identical to a conscious person — same brain states, same behavior, same dispositions, same verbal reports — but lacking any phenomenal consciousness whatsoever. The zombie sees red things, processes the visual input, says "that's red," uses red-vocabulary correctly, and yet there is *nothing it is like* to be the zombie having those visual states. The lights are off behind the functional façade.

The Thought Experiment

The thought experiment, in its sharpest form, comes from David Chalmers (especially in *The Conscious Mind*, 1996). Zombies are *conceivable*, Chalmers argues: we can coherently imagine them, the scenario generates no contradictions, the description stays internally consistent. From the conceivability of zombies, Chalmers argues to their metaphysical *possibility*. From their

possibility, he argues that phenomenal properties cannot be necessitated by physical properties. From that, he concludes that physicalism is false — and by extension, that any naturalism strong enough to entail physicalism is false.

The argument ranks among the most influential in contemporary philosophy of mind.

The Argument's Structure

In compact form, Chalmers's zombie argument has three steps.

1. **Conceivability premise.** Philosophical zombies are conceivable. We can coherently imagine a physical and functional duplicate of a conscious being that lacks phenomenal consciousness.
2. **Conceivability-to-possibility bridge.** What is ideally conceivable is metaphysically possible. (This is the controversial bridge premise. Different versions of the argument argue for it in different ways.)
3. **Possibility-to-anti-physicalism.** If zombies are metaphysically possible, then phenomenal properties are not necessitated by physical properties; that is, physicalism is false.

Each step does real work, and physicalist replies engage all three.

Three Lines of Reply

Line 1: The conceivability premise may depend on the inner-theater picture.

When we "imagine" a zombie, we hold fixed all the physical and functional facts and then *subtract* the phenomenal facts. This subtraction operation only seems coherent, the intentionalist argues, if phenomenal facts are conceived as something *over and above* the physical and functional facts — an additional inner property that could be present or absent without changing the underlying physical-functional structure.

But that conception of phenomenal facts is exactly the inner-theater picture, the very picture strong representationalism argues against. Once phenomenal character is recast as *how* a system represents the world, the subtraction operation loses its target. There is no inner item to subtract while keeping everything else the same; phenomenal character is internal to the representational activity, not a free-floating addition to it.

So, this reply concludes, zombies are not conceivable in the sense the argument requires. We can imagine *something* — a robot, a behavioral mimic — but we cannot coherently imagine a being identical in every physical-functional respect that lacks phenomenal consciousness, because there is no phenomenal item left to subtract from the physical-functional structure that constitutes it. This line targets the conceivability premise itself; accepting ideal conceivability means declining it as a complete answer.

Line 2: The conceivability-to-possibility bridge has contested scope.

Even granting that something coherently imaginable counts as "conceivable," the move from conceivability to metaphysical possibility requires defense. Many apparent counterexamples involve water without H_2O, gold without atomic number 79, or other identities discovered *a posteriori*. Chalmers built his reply around exactly that problem. His bridge concerns *ideal primary positive*

conceivability. Roughly, a scenario must remain coherent under ideal reflection when considered as an account of how the actual world might have turned out. Ordinary cases can preserve a string of words or a secondary imagining while losing their appearance of genuine possibility. The live dispute concerns whether ideal primary conceivability reaches metaphysical possibility in the phenomenal case. Casual imagining settles nothing.

Line 3: The type-B dispute turns on strong necessity.
If a complete physical duplicate without consciousness is metaphysically possible, physicalism is false. Type-B physicalism has no room to grant that step while keeping its defining claim that the physical facts necessitate the phenomenal facts. It grants the epistemic premise instead: the zombie remains ideally conceivable because phenomenal and physical concepts reach one state by independent routes.

Chalmers argues that this reply leaves type-B physicalism with a *strong necessity*. Ordinary *a posteriori* identities involve a contingent way of fixing reference: "water" initially picks out the watery stuff around us, and science discovers that stuff to be H_2O. Core phenomenal concepts, he argues, have no comparable reference-fixer waiting behind the felt character. Their primary and secondary intensions coincide. If the zombie remains ideally conceivable under those concepts, the conceiving carries modal weight. The phenomenal-concept strategy explains the epistemic separation, but it does not by itself establish the identity or show that a strong necessity holds. The remaining dispute concerns the governing modal rule. Chalmers accepts it; type-B physicalists deny that every ideal conceptual separation describes a possible world.

The Conceivability Trap
One intentionalist diagnosis treats the zombie argument as a trap. The imagination operation — hold the physical fixed, subtract the phenomenal — feels straightforward because the inner-theater picture has already cast phenomenal character as an extra item. That diagnosis explains some of the thought experiment's grip. It does not, by itself, show that ideal zombie conceivability fails. Chalmers's modal argument was built to survive the observation that ordinary imagining can smuggle in a description; treating the diagnosis as a refutation would simply restate Line 1.

Where the Argument Still Earns Respect
The zombie argument repays serious engagement even from those who reject its conclusion. Three things it gets right:

The pull toward dualism is real. Many thoughtful readers find the zombie argument compelling because phenomenal consciousness seems separable in some way from the physical-functional facts. The inner-theater diagnosis explains part of that pull. The persistence of ideal conceivability after that picture has been rejected requires the further phenomenal-concept and modal story.

The explanatory gap is real. Even for a reader who denies the argument its conclusion, it points at something: third-person physical descriptions of brain states do not generate the phenomenal concepts that first-person experience produces. That gap is real, and Sellars's Myth-of-the-

Given diagnosis together with the phenomenal-concept strategy supply a handle on it. See *The Explanatory Gap*.

The hard problem deserves articulation. Chalmers's distinction between the easy problems and the hard problem works as a framing tool even for physicalists who treat the gap as epistemic and reject its ontological conclusion. See *The Hard Problem of Consciousness*.

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

- Scott Aaronson. "Why I Am Not an Integrated Information Theorist (or, the Unconscious Expander)." *Shtetl-Optimized* (blog), May 2014. App. B n.62
- Robert M. Adams. *Leibniz: Determinist, Theist, Idealist* (Oxford: Oxford University Press, 1994). Ch12 n.8
- 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
- J. L. Austin. *How to Do Things with Words*, ed. J. O. Urmson (Oxford: Clarendon Press, 1962). Ch16 n.11
- J. L. Austin. *Sense and Sensibilia*, reconstructed from manuscript notes by G. J. Warnock (Oxford: Oxford University Press, 1962). Ch03 n.3, Argument from Illusion n.4, Argument from Illusion 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* (London: Routledge, 2001). Ch19 n.1
- A. J. Ayer. *The Foundations of Empirical Knowledge* (London: Macmillan, 1940). Ch03 n.1, Argument from Illusion n.10
- Lynne Rudder Baker. *Persons and Bodies: A Constitution View* (Cambridge: Cambridge University Press, 2000). Ch06 n.21
- Justin L. Barrett. *Why Would Anyone Believe in God?* (Walnut Creek, CA: AltaMira Press, 2004). Ch25 n.7
- Friedrich Beck and John Eccles. "Quantum Aspects of Brain Activity and the Role of Consciousness." *Proceedings of the National Academy of Sciences* 89 (1992). Ch09 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). Ch23 n.1, Ch24 n.1, Ch25 n.1, App. B n.4
- Jonathan Bennett. *Locke, Berkeley, Hume: Central Themes* (Oxford: Clarendon Press, 1971). Ch02 n.7, Ch02 n.8, Ch02 n.9
- José Luis Bermúdez. *Thinking Without Words* (Oxford: Oxford University Press, 2003). Ch17 n.2
- Max Black. "The Identity of Indiscernibles." *Mind* 61 (1952). Ch12 n.3
- Ned Block. "Advertisement for a Semantics for Psychology." *Midwest Studies in Philosophy* 10 (1986). Ch20 n.11, Ch23 n.4
- Ned Block. "Inverted Earth." *Philosophical Perspectives* 4 (1990). Ch02 n.5, Ch06 n.15, Inverted Earth n.1

- 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 (1995). Ch24 n.13
- Ned Block. “Psychologism and Behaviorism.” *Philosophical Review* 90 (1981). Ch24 n.1
- Ned Block. “The Higher Order Approach to Consciousness Is Defunct.” *Analysis* 71, no. 3 (2011): 419–431. Ch06 n.7b, App. B n.58
- Gemma Boleda. “Distributional Semantics and Linguistic Theory.” *Annual Review of Linguistics* 6 (2020). 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. Transparency nn.4, 10
- 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
- Bill Brewer. *Perception and Its Objects* (Oxford: Oxford University Press, 2011). Argument from Illusion n.6, Argument from Illusion n.9
- C. D. Broad. *Scientific Thought* (London: Routledge & Kegan Paul, 1923). Ch02 n.10
- Tyler Burge. “Individualism and the Mental.” *Midwest Studies in Philosophy* 4 (1979): 73–121. DOI 10.1111/j.1475-4975.1979.tb00374.x. Ch16 n.15, App. B n.4
- Tyler Burge. *Origins of Objectivity* (Oxford: Oxford University Press, 2010). Ch13 n.6, Ch14 n.1
- Patrick Butlin et al. “Identifying Indicators of Consciousness in AI Systems.” *Trends in Cognitive Sciences* (2025). Ch24 n.6, Coda n.5
- Patrick Butlin et al. *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness* (2023). Ch24 n.6
- Alex Byrne. “Intentionalism Defended.” *Philosophical Review* 110 (2001). Ch04 n.3, Ch06 n.4, Transparency n.8
- Alex Byrne and Heather Logue, eds. *Disjunctivism: Contemporary Readings* (Cambridge, MA: MIT Press, 2009). Ch03 n.4, Argument from Hallucination n.11
- Alex Byrne and Michael Tye. “Qualia ain’t in the Head.” *Noûs* 40 (2006). Ch15 n.7

- 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. Ch22 n.15, Ch24 n.12, Coda n.7, App. B n.61
- David J. Chalmers. "Does Conceivability Entail Possibility?" *Conceivability and Possibility* (Oxford: Oxford University Press, 2002). Ch09 n.5, App. B n.12, App. B n.13
- David J. Chalmers. "Facing Up to the Problem of Consciousness." *Journal of Consciousness Studies* 2 (1995). Ch06 n.22, Ch09 n.3
- David J. Chalmers. "Phenomenal Concepts and the Explanatory Gap." *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism* (New York: Oxford University Press, 2007). Ch08 n.14, Ch11 n.8, App. B n.11, App. B n.13, App. B n.17, App. B n.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 n.15, App. B n.17, App. B n.55
- David J. Chalmers. "The Two-Dimensional Argument Against Materialism." *The Character of Consciousness* (New York: Oxford University Press, 2010). 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). Ch02 n.2, Ch06 n.22, Ch09 n.3, Ch11 n.3, Ch20 n.18, Ch22 n.5, App. B n.9, App. B n.32
- Andy Clark. *Being There: Putting Brain, Body, and World Together Again* (Cambridge, MA: MIT Press, 1997). Ch15 n.6
- Jonathan Cohen. *The Red and the Real: An Essay on Color Ontology* (Oxford: Oxford University Press, 2009). App. B n.23
- 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
- David Cole. "Artificial Intelligence and Personal Identity." *Synthese* 88 (1991). Ch21 n.3
- Clara Colombatto and Stephen M. Fleming. "Folk Psychological Attributions of Consciousness to Large Language Models." *Neuroscience of Consciousness* (2024). Ch25 n.4
- B. Jack Copeland. "The Church–Turing Thesis." *Stanford Encyclopedia of Philosophy* (2000). App. B n.18
- Luca Corti. "Organizational Normativity and Teleology: A Critique." *Synthese* 202 (2023). Ch24 n.3
- Tim Crane. "Introspection, Intentionality, and the Transparency of Experience." *Philosophical Topics* 28 (2000). Ch04 n.7, Ch19 n.3

- Tim Crane. “Is There a Perceptual Relation?” In *Perceptual Experience*, ed. Tamar Szabó Gendler and John Hawthorne (Oxford: Oxford University Press, 2006), 126–146. Ch02 n.5, Ch04 n.7, Ch05 nn.4, 8, Ch14 n.13, Ch23 n.13, Argument from Illusion nn.5, 12, Argument from Hallucination nn.8, 12
- Tim Crane. “What Is the Problem of Perception?” *Synthesis Philosophica* 20, no. 2 (2005): 237–264. Argument from Hallucination n.8
- Tim Crane. *The Objects of Thought* (Oxford: Oxford University Press, 2013). Ch01 n.3, App. B n.28, App. B n.29
- Tim Crane and Craig French. “The Problem of Perception.” *Stanford Encyclopedia of Philosophy* (Fall 2021 edition). Ch05 nn.4, 8, Ch06 n.4, Ch23 n.13, Argument from Hallucination n.12
- Brian Cutter and Michael Tye. “Tracking Representationalism and the Painfulness of Pain.” *Philosophical Issues* 21 (2011). 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
- 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. *Brainchildren: Essays on Designing Minds* (Cambridge, MA: MIT Press, 1998). Ch24 n.2
- Daniel C. Dennett. *Consciousness Explained* (Boston: Little, Brown, 1991). Ch17 n.9, Ch24 n.5
- Daniel C. Dennett. *Darwin’s Dangerous Idea* (New York: Simon & Schuster, 1995). Ch24 n.5
- Daniel C. Dennett. *Elbow Room: The Varieties of Free Will Worth Wanting* (Cambridge, MA: MIT Press, 1984). Ch18 n.8
- Daniel C. Dennett. *From Bacteria to Bach and Back: The Evolution of Minds* (New York: W. W. Norton, 2017). Ch24 n.1
- Daniel C. Dennett. *The Intentional Stance* (Cambridge, MA: MIT Press, 1987). Ch01 n.4, Ch24 n.1, Ch24 n.5, Ch25 n.12
- David Deutsch. “Quantum Theory, the Church–Turing Principle and the Universal Quantum Computer.” *Proceedings of the Royal Society of London A* 400 (1985). Ch22 n.11, App. B n.21
- Fred Dretske. *Explaining Behavior: Reasons in a World of Causes* (Cambridge, MA: MIT Press, 1988). Ch15 n.1, Ch15 n.5, Ch18 n.9, Ch22 n.6, Ch22 n.12, Ch23 n.11, Inverted Earth n.13
- Fred Dretske. *Knowledge and the Flow of Information* (Cambridge, MA: MIT Press, 1981). Ch05 n.3, Ch06 n.19, Ch13 n.1, Ch13 n.6, Ch14 n.2, Ch14 n.10, Ch15 n.1, App. B n.7

- Fred Dretske. *Naturalizing the Mind* (Cambridge, MA: MIT Press, 1995). Ch05 n.3, Ch06 n.19, Ch09 n.4, Ch13 n.6, Ch15 n.5, Ch15 n.7, Ch20 n.17, Ch22 n.12, App. B n.38, App. B n.48
- Hubert L. Dreyfus. *What Computers Can't Do* (New York: Harper & Row, 1972). Ch20 n.21
- Hubert L. Dreyfus. *What Computers Still Can't Do: A Critique of Artificial Reason* (Cambridge, MA: MIT Press, 1992). Ch20 n.21
- Leonard Dung and Luke Kersten. "Implementing Artificial Consciousness." *Mind & Language* 40 (2025). Ch24 n.8
- 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
- J. R. Firth. "A Synopsis of Linguistic Theory, 1930–1955." *Studies in Linguistic Analysis* (1957). 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.8
- 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.6
- Jerry A. Fodor. *A Theory of Content and Other Essays* (Cambridge, MA: MIT Press, 1990). Ch15 n.5, App. B n.49
- Jerry A. Fodor. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (Cambridge, MA: MIT Press, 1987). Ch20 n.11, 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). Ch03 n.4
- 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 (2010). Ch22 n.16, Ch24 n.6
- 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
- Philip Goff. *Consciousness and Fundamental Reality* (Oxford: 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
- Mitchell Green. "Speech Acts." *Stanford Encyclopedia of Philosophy*, substantive revision September 24, 2020. https://plato.stanford.edu/entries/speech-acts/. Ch16 n.1
- 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. Ch16 n.9
- 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. Ch23 n.10
- 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). Ch04 n.1, Ch05 n.1, Ch06 n.2, Transparency n.1, App. B n.14
- Stevan Harnad. "The Symbol Grounding Problem." *Physica D: Nonlinear Phenomena* 42 (1990). Ch13 n.2, Ch23 n.3, Ch24 n.10
- Sam Harris. *Free Will* (New York: Free Press, 2012). Ch18 n.2
- Zellig S. Harris. "Distributional Structure." *Word* 10 (1954). Ch13 n.8
- Vladimír Havlík. "Meaning and Understanding in Large Language Models." *Synthese* 205 (2025). Ch22 n.17
- John Hawthorne and Karson Kovakovich. "Disjunctivism." *Proceedings of the Aristotelian Society, Supplementary Volumes* 80 (2006): 145–183. Argument from Hallucination n.12
- Martin Heidegger. *Being and Time* (1927). Ch20 n.21
- 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

- David R. Hilbert. *Color and Color Perception: A Study in Anthropocentric Realism* (Stanford: CSLI, 1987). App. B n.25
- Christopher Hill. *Sensations: A Defense of Type Physicalism* (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
- J. M. Hinton. "Visual Experiences." *Mind* 76 (1967). Ch05 n.7
- J. M. Hinton. *Experiences: An Inquiry into Some Ambiguities* (Oxford: Clarendon Press, 1973). Ch05 n.7
- Jakob Hohwy. *The Predictive Mind* (Oxford: Oxford University Press, 2013). Ch02 n.11
- 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
- Martin Hägglund. *This Life: Secular Faith and Spiritual Freedom* (New York: Pantheon, 2019). Ch24 n.9
- Peter van Inwagen. *An Essay on Free Will* (Oxford: Clarendon Press, 1983). App. B n.65
- Frank Jackson. "Epiphenomenal Qualia." *Philosophical Quarterly* 32 (1982). Ch06 n.12, Ch11 n.7
- Frank Jackson. "What Mary Didn't Know." *Journal of Philosophy* 83 (1986). Ch06 n.12, Ch11 n.7
- Frank Jackson. *Perception: A Representative Theory* (Cambridge: Cambridge University Press, 1977). Ch03 n.2
- Pierre Jacob. "Intentionality." *Stanford Encyclopedia of Philosophy* (2019). Ch24 n.5
- Hilla Jacobson. "Killing the Messenger: Representationalism and the Painfulness of Pain." *Philosophical Quarterly* 63 (2013). Ch07 n.5
- 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 n.3, 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
- François Kammerer. "Can You Believe It? Illusionism and the Illusion Meta-Problem." *Philosophical Psychology* 31, no. 1 (2018). App. B n.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. App. B n.56
- Robert Kane. *The Significance of Free Will* (New York: Oxford University Press, 1996). App. B n.65
- 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* (2026). Ch25 n.4
- Jaegwon Kim. *Mind in a Physical World* (Cambridge, MA: MIT Press, 1998). Ch09 n.7, Ch18 n.9
- Matthew Kinakin. "Why Intentionalists Can't Take Painkillers." *Philosophical Studies* 183 (2026). Ch07 n.5
- Amy Kind. "What's So Transparent About Transparency?" *Philosophical Studies* 115 (2003). Ch04 n.4, Ch06 n.3, Ch06 n.9, Ch11 n.19, Transparency n.4
- Colin Klein. *What the Body Commands: The Imperative Theory of Pain* (Cambridge, MA: MIT Press, 2015). Ch07 n.5
- Kepa Korta and John Perry. "Pragmatics." *Stanford Encyclopedia of Philosophy*, substantive revision May 28, 2024. https://plato.stanford.edu/entries/pragmatics/. Ch16 n.1
- Uriah Kriegel. *Subjective Consciousness: A Self-Representational Theory* (Oxford: Oxford University Press, 2009). Ch06 n.7b, App. B n.59
- Uriah Kriegel. *The Routledge Handbook of Franz Brentano and the Brentano School* (London: Routledge, 2017). App. B n.28
- Saul Kripke. *Naming and Necessity* (Cambridge, MA: Harvard University Press, 1980). Ch11 n.5, Ch16 n.13
- Saul Kripke. *Wittgenstein on Rules and Private Language* (Cambridge, MA: Harvard University Press, 1982). Ch17 n.5
- Ray Kurzweil. *How to Create a Mind* (New York: Viking, 2012). Ch24 n.7
- Ray Kurzweil. *The Singularity Is Near: When Humans Transcend Biology* (New York: Viking, 2005). Ch24 n.7, Coda n.1
- Peter Langland-Hassan. "Inner Speech." *Stanford Encyclopedia of Philosophy* (Fall 2023 edition). Ch17 n.1
- 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
- Alessandro Lenci. "Distributional Models of Word Meaning." *Annual Review of Linguistics* 4 (2018). Ch23 n.5
- Joseph Levine. "Materialism and Qualia: The Explanatory Gap." *Pacific Philosophical Quarterly* 64 (1983): 354–361. Ch02 n.3, Ch09 n.2, Ch11 n.1, App. B n.10

- Joseph Levine. *Purple Haze: The Puzzle of Consciousness* (New York: Oxford University Press, 2001). Ch02 n.3, Ch11 n.1
- David Lewis. "Are We Free to Break the Laws?" *Theoria* 47, no. 3 (1981): 113–121. App. B n.65
- 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
- Benjamin Libet. "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action." *Behavioral and Brain Sciences* 8, no. 4 (1985): 529–566. 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.12, Ch08 n.14, Ch09 n.5, Ch11 n.6, App. B n.16, App. B n.17
- 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. "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* (2019). Ch06 n.20
- William G. Lycan. "The Intentionality of Smell." *Frontiers in Psychology* 5 (2014). Ch07 n.8
- William G. Lycan. *Consciousness and Experience* (Cambridge, MA: MIT Press, 1996). Inverted Earth n.17, App. B n.42
- Fintan Mallory. "Teleosemantics for Neural Word Embeddings." *Mind & Language* online (2026). Ch14 n.11, Ch23 n.7, Ch24 n.4, App. B n.51
- 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. Ch16 n.18, Ch23 n.12, Ch24 n.10
- M. G. F. Martin. "The Limits of Self-Awareness." *Philosophical Studies* 120 (2004). 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 and Language* 17 (2002). Ch03 n.4, Ch23 n.14
- 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. App. B n.60

- John McDowell. "Criteria, Defeasibility, and Knowledge." *Proceedings of the British Academy* 68 (1982). Ch03 n.3, Ch05 n.7, Ch19 n.2
- John McDowell. *Mind and World* (Cambridge, MA: Harvard University Press, 1994). Ch03 n.4, Ch06 n.20, Argument from Hallucination n.10, App. B n.1
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021). Ch21 n.6, Ch24 n.1
- Angela Mendelovici. "Reliable Misrepresentation and Tracking Theories of Mental Representation." *Philosophical Studies* 165 (2013). Ch13 n.1
- Angela Mendelovici. *The Phenomenal Basis of Intentionality* (New York: Oxford University Press, 2018). Ch07 n.9, Ch09 n.4, Ch13 n.1
- Maurice Merleau-Ponty. *Phenomenology of Perception* (1945). Ch20 n.21
- Ruth Garrett Millikan. "Biosemantics." *Journal of Philosophy* 86 (1989). Ch14 n.6, Ch14 n.9, Ch15 n.4
- Ruth Garrett Millikan. "What Has Natural Information to Do with Intentional Representation?" *Royal Institute of Philosophy Supplement* 49 (2001). Ch14 n.12
- Ruth Garrett Millikan. *Language, Thought, and Other Biological Categories: New Foundations for Realism* (Cambridge, MA: MIT Press, 1984). Ch06 n.19, Ch14 n.6, Ch14 n.9, Ch15 n.2, Ch15 n.4, Ch22 n.6, Ch23 n.11, Ch24 n.4, App. B n.38
- Ruth Garrett Millikan. *White Queen Psychology and Other Essays for Alice* (Cambridge, MA: MIT Press, 1993). Ch15 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). Ch21 n.7
- Kevin J. Mitchell. "The Origins of Meaning: From Pragmatic Control Signals to Semantic Representations." *Philosophy and the Mind Sciences* 7 (2026). Ch15 n.4
- Maël Montévil and Matteo Mossio. "Biological Organisation as Closure of Constraints." *Journal of Theoretical Biology* 372 (2015). Ch24 n.3
- G. E. Moore. "The Refutation of Idealism." *Mind* 12 (1903). Ch04 n.4, Ch05 n.1, Ch06 n.3, Inverted Earth n.11, Transparency n.2
- Peter Ludlow, Yujin Nagasawa, and Daniel Stoljar. *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
- Thomas Nagel. *The View from Nowhere* (New York: Oxford University Press, 1986). Ch09 n.1
- Karen Neander. *A Mark of the Mental: In Defense of Informational Teleosemantics* (Cambridge, MA: MIT Press, 2017). Ch14 n.5, Ch23 n.15, App. B n.46

- 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
- Shaun Nichols and Joshua Knobe. “Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions.” *Noûs* 41, no. 4 (2007): 663–685. App. B n.66
- Martine Nida-Rümelin and Donnchadh Ó Conaill. “Qualia: The Knowledge Argument.” *Stanford Encyclopedia of Philosophy* (substantive revision March 1, 2024). Ch08 n.3, Ch08 n.11
- Ezequiel Di Paolo. “Autopoiesis, Adaptivity, Teleology, Agency.” *Phenomenology and the Cognitive Sciences* 4 (2005). Ch24 n.3
- David Papineau. “Mind the Gap.” *Philosophical Perspectives* 12 (1998). Ch11 n.8
- David Papineau. *Philosophical Naturalism* (Oxford: Blackwell, 1993). Ch11 n.8
- David Papineau. *Thinking About Consciousness* (Oxford: Clarendon Press, 2002). 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.7
- Roger Penrose. *The Emperor’s New Mind* (Oxford: Oxford University Press, 1989). App. B n.21
- Derk Pereboom. *Living Without Free Will* (Cambridge: Cambridge University Press, 2001). Ch18 n.2
- John Perry. *Knowledge, Possibility, and Consciousness* (Cambridge, MA: MIT Press, 2001). Ch08 n.11
- Steven T. Piantadosi and Felix Hill. “Meaning without reference in large language models.” arXiv:2208.02957 (2022). Ch16 n.17
- 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 (2007). App. B n.20
- 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. https://plato.stanford.edu/entries/compositionality/. Ch16 n.4
- H. H. Price. *Perception* (London: Methuen, 1932). Ch03 n.1, Argument from Illusion n.1, Argument from Illusion n.10
- 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, Ch22 n.18

- Hilary Putnam. *Representation and Reality* (Cambridge, MA: MIT Press, 1988). Ch17 n.1, Ch20 n.8, Ch20 n.9, Ch20 n.10, Ch20 n.11, Ch22 n.1
- V. S. Ramachandran and Sandra Blakeslee. *Phantoms in the Brain: Probing the Mysteries of the Human Mind* (New York: William Morrow, 1998). Ch07 n.1
- Matthew Ratcliffe. *Feelings of Being: Phenomenology, Psychiatry and the Sense of Reality* (Oxford: Oxford University Press, 2008). Ch07 n.10
- Vasudevi Reddy. *How Infants Know Minds* (Cambridge, MA: 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* (Cambridge: Cambridge University Press, 1996). Ch25 n.5
- Indrek Reiland. "The Unity of Perceptual Content." *The Philosophical Quarterly* 74, no. 3 (2024): 941–961. Argument from Hallucination n.5, Transparency n.10
- Howard Robinson. *Perception* (London: Routledge, 1994). Ch02 n.9, Ch02 n.10, Ch03 n.1, Ch03 n.2, Ch05 n.4, Argument from Illusion n.3, Argument from Illusion n.8, Argument from Illusion n.11
- David M. Rosenthal. *Consciousness and Mind* (Oxford: Clarendon Press, 2005). Ch06 n.7b, App. B n.57
- David Rosenthal. "Exaggerated Reports: Reply to Block." *Analysis* 71, no. 3 (2011): 431–437. Ch06 n.7b, App. B n.58
- Bertrand Russell. "On Denoting." *Mind*, n.s. 14, no. 56 (1905): 479–493. DOI 10.1093/mind/XIV.4.479. Ch16 n.3
- Bertrand Russell. "Knowledge by Acquaintance and Knowledge by Description." *Proceedings of the Aristotelian Society* 11 (1910). Ch12 n.9
- Bertrand Russell. *The Analysis of Matter* (London: Kegan Paul, 1927). Ch10 n.1
- Bertrand Russell. *The Problems of Philosophy* (London: Williams and Norgate, 1912). Ch02 n.10
- 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
- Roger Schank and Robert Abelson. *Scripts, Plans, Goals and Understanding* (Hillsdale, NJ: Lawrence Erlbaum, 1977). Ch21 n.1
- Marya Schechtman. *The Constitution of Selves* (Ithaca: Cornell University Press, 1996). Ch17 n.7
- Susanna Schellenberg. *The Unity of Perception: Content, Consciousness, Evidence* (Oxford: Oxford University Press, 2018). Argument from Hallucination nn.5, 7

- Eric Schwitzgebel. "Borderline Consciousness, When It's Neither Determinately True nor Determinately False That Experience Is Present." *Philosophical Studies* (2023). Ch10 n.7
- John R. Searle. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (1980). Ch14 n.13, Ch21 n.1, Ch22 n.8, Ch24 n.1
- John R. Searle. *Intentionality: An Essay in the Philosophy of Mind* (Cambridge: Cambridge University Press, 1983). Ch01 n.2, Ch01 n.4, 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, Ch24 n.1
- 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 n.2, Ch22 n.9, Ch24 n.11, App. B n.6
- John R. Searle. *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992). Ch01 n.4, Ch11 n.15, Ch13 n.3, Ch13 n.7, Ch21 n.5, Ch24 n.5, App. B n.30
- Wilfrid Sellars. *Science, Perception and Reality* (London: Routledge & Kegan Paul, 1963). Ch17 n.6
- Anil Seth. *Being You: A New Science of Consciousness* (London: Faber & Faber, 2021). Ch02 n.11
- Murray Shanahan. "Talking about Large Language Models." *Communications of the ACM* 67, no. 2 (2024): 68–79. Ch21 n.6
- Claude E. Shannon. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27 (1948). Ch13 n.5
- Nicholas Shea. "Realist Representational Explanations of Agency Should Require Some Unity of Purpose." *Philosophy and the Mind Sciences* 7 (2026). Ch24 n.4, App. B n.51
- Nicholas Shea. *Representation in Cognitive Science* (Oxford: Oxford University Press, 2018). Ch15 n.4, Ch24 n.4, App. B n.50
- Sydney Shoemaker. "Self-Knowledge and 'Inner Sense'." *Philosophy and Phenomenological Research* 54 (1994). Ch04 n.2
- Susanna Siegel. *The Contents of Visual Experience* (Oxford: Oxford University Press, 2010). Argument from Illusion n.7
- Jonathan Simon. "The Hard Problem of the Many." *Philosophical Perspectives* 31 (2017). Ch10 n.6
- 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. App. B n.63

- A. D. Smith. *The Problem of Perception* (Cambridge, MA: Harvard University Press, 2002). Ch05 n.4, Ch05 n.8, Argument from Illusion n.12, Argument from Illusion n.13
- Brian Cantwell Smith. *The Promise of Artificial Intelligence: Reckoning and Judgment* (Cambridge, MA: MIT Press, 2019). Ch24 n.1
- Paul Snowdon. "Perception, Vision and Causation." *Proceedings of the Aristotelian Society* 81 (1981). Ch03 n.4, Ch05 n.7
- Paul Snowdon. "The Formulation of Disjunctivism: A Response to Fish." *Proceedings of the Aristotelian Society* 105 (2005). Ch06 n.18
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (2014). Ch23 n.6
- 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). Ch22 n.5
- Daniel Stoljar. "The Argument from Diaphanousness." In *Language, Mind and World*, *Canadian Journal of Philosophy*, supplementary vol. 30 (2004): 341–390. Ch04 n.4, Transparency n.4
- Galen Strawson. "Against Narrativity." *Real Materialism and Other Essays* (Oxford: Clarendon Press, 2008). Ch17 n.13
- Galen Strawson. "Real Materialism." *Real Materialism and Other Essays* (Oxford: Clarendon Press, 2008). Ch06 n.6
- P. F. Strawson. "Freedom and Resentment." *Proceedings of the British Academy* 48 (1962): 1–25. Ch17 n.12, Ch18 n.7
- P. F. Strawson. *Individuals: An Essay in Descriptive Metaphysics* (London: Methuen, 1959). Ch12 n.4
- P. F. Strawson. *The Bounds of Sense: An Essay on Kant's "Critique of Pure Reason"* (London: Methuen, 1966). Ch12 n.13
- Evan Thompson. *Mind in Life* (Cambridge, MA: Harvard University Press, 2007). Ch24 n.3, App. B n.39
- Francisco Varela, Evan Thompson, and Eleanor Rosch. *The Embodied Mind: Cognitive Science and Human Experience* (Cambridge, MA: MIT Press, 1991). Ch15 n.6, Ch21 n.4
- 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
- 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. App. B n.63
- Endel Tulving. *Elements of Episodic Memory* (Oxford: Oxford University Press, 1983). Ch02 n.13
- Alan Turing. "Computing Machinery and Intelligence." *Mind* 59 (1950). Ch24 n.2
- 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
- Michael Tye. "How I Learned to Stop Worrying and Love Panpsychism." *Journal of Consciousness Studies* 31, nos. 9–10 (2024): 10–28. Ch10 nn.5–7, 10, Inverted Earth n.20, App. B n.40
- Michael Tye. "Intentionalism and the Argument from No Common Content." *Philosophical Perspectives* 21 (2007): 589–613. App. B n.14
- Michael Tye. "Phenomenal Consciousness: The Explanatory Gap as a Cognitive Illusion." *Mind* 108 (1999). Ch11 n.8
- Michael Tye. "Representationalism and the Transparency of Experience." *Noûs* 36 (2002). Transparency n.8
- Michael Tye. *Consciousness Revisited: Materialism Without Phenomenal Concepts* (Cambridge, MA: MIT Press, 2009). Ch08 n.9, Ch11 n.8
- Michael Tye. *Consciousness, Color, and Content* (Cambridge, MA: MIT Press, 2000). Ch02 n.13, Ch04 n.6, Ch06 n.7, Ch06 n.10, Ch06 n.21, Ch07 n.10, Ch11 n.8, Ch14 n.7, Inverted Earth n.6, Inverted Earth n.8, Inverted Earth n.16, Transparency n.8, App. B n.14, App. B n.25, App. B n.38, App. B n.45
- Michael Tye. *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind* (Cambridge, MA: MIT Press, 1995). Ch04 n.6, Ch06 n.4, Ch07 n.2, Ch07 n.10, Ch09 n.4, Ch20 n.17, Ch22 n.12, Ch24 n.13, App. B n.14, App. B n.25, App. B n.27, App. B n.34
- Michael Tye. *Vagueness and the Evolution of Consciousness: Through the Looking Glass* (New York: Oxford University Press, 2021). Ch02 n.4, Ch10 nn.5, 7, Inverted Earth n.20, App. B nn.34, 40
- J. David Velleman. "The Self as Narrator." *Self to Self: Selected Essays* (Cambridge: Cambridge University Press, 2006). Ch17 n.10
- Lev Vygotsky. *Thought and Language* (1934; trans. Eugenia Hanfmann and Gertrude Vakar, Cambridge, MA: MIT Press, 1962). App. B n.1
- Zina B. Ward. "Directive Representation and the Job Description Challenge." *Philosophy and the Mind Sciences* 7 (2026). Ch15 n.4

- Jason Wei et al. “Emergent Abilities of Large Language Models.” *Transactions on Machine Learning Research* (2022). Ch21 n.7
- 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.23
- Bernard Williams. “The Makropulos Case: Reflections on the Tedium of Immortality.” *Problems of the Self* (Cambridge: Cambridge University Press, 1973). Ch24 n.9
- Iwan Williams. “Can Structural Correspondences Ground Real-World Representational Content in Large Language Models?” *Mind & Language* (2026). Ch23 n.8
- Ludwig Wittgenstein. *Philosophical Investigations*, trans. G. E. M. Anscombe, 3rd ed. (Oxford: Basil Blackwell, 1967; English-text reprint, 1972). Ch16 nn.7–8
- Helen Yetter-Chappell. “Dissolving Type-B Physicalism.” *Philosophical Perspectives* 31 (2017). Ch11 n.18
- Helen Yetter-Chappell. “Guessing at Ghosts in the Machine.” *Ratio* 39 (2026). Ch25 n.11
- John W. Yolton. *Perception and Reality: A History from Descartes to Kant* (Ithaca: Cornell University Press, 1996). Ch02 n.8
- John W. Yolton. *Perceptual Acquaintance from Descartes to Reid* (Minneapolis: University of Minnesota Press, 1984). Ch02 n.8
- Dan Zahavi. “Beyond Empathy: Phenomenological Approaches to Intersubjectivity.” *Journal of Consciousness Studies* 8 (2001). Ch19 n.3