AI Consciousness Requires Validated Models of Human Consciousness

Paras Chopra
Lossfunk
paras@lossfunk.com

Abstract

Debates about AI consciousness often proceed without grounding the concept in empirically validated models. This position paper argues that meaningful claims about AI consciousness should be licensed by (and graded by confidence in) models validated on humans. Drawing on Quine's observation sentences and pragmatic philosophy of science, we argue that all scientific observation ultimately depends on human perceptual agreement, including observations about consciousness itself. Without validated human models that make testable predictions about conscious experience, the question "Is this AI conscious?" lacks sufficient empirical grounding necessary for scientific progress. We propose a human-first methodology: identify measurable phenomena associated with consciousness in humans, build predictive models, validate them empirically, and only then apply these models to AI systems. This approach accelerates philosophical debates into productive scientific inquiry.

Introduction

As AI systems grow increasingly sophisticated, questions about their potential consciousness have moved from philosophical speculation to urgent practical concern (Butlin, Long et al. 2023; Schwitzgebel 2025). Could large language models have subjective experiences? Do reinforcement learning agents feel anything? These questions carry significant implications for AI ethics, safety, and policy.

But current debates about AI consciousness suffer from a fundamental problem. The concept of "consciousness" itself lacks the empirical grounding necessary for scientific adjudication. As Schwitzgebel (Schwitzgebel 2025) argues, we will soon create AI systems that are conscious according to some mainstream theories but not others, and we will have no principled way to determine which theories are correct. This is not merely a temporary gap in our knowledge; it reflects a deeper methodological problem.

Our position is this: **meaningful claims about AI consciousness should be licensed by (and graded by confidence in) by models validated on humans**. Without such models, the term "consciousness" cannot be properly grounded, and debates about machine consciousness will remain unproductive in precisely the sense that pragmatist philosophers warned against (James 1907; Dewey 1929). Our methodology is pragmatically motivated as it essentially operationalizes consciousness as its measurable correlates in humans. While we acknowledge the hard problem (Chalmers 1995), we align with Seth's 'real problem' approach (Seth 2016) and Dennett's contention (Dennett 1996) that the hard problem may dissolve in the process of solving easy problems not through philosophical argument, but through accumulating successful explanations.

This paper develops three interconnected arguments. First, we diagnose the grounding problem: "consciousness" is a family-resemblance concept that admits multiple meanings, and without specifying which aspect we mean and how to measure it, claims about AI consciousness lack empirical content (Section 2). Second, we argue that all scientific observation, including observations relevant to consciousness, ultimately depends on human perceptual agreement, making human consciousness the necessary starting point for any empirical program (Section 3). Third, we propose a human-first methodology that can transform metaphysical confusion into productive research (Section 4).

Our approach builds on but is distinct from recent empirical programs in consciousness science. Butlin et al. (Butlin, Long et al. 2023) derive indicator properties from existing theories and assess AI systems against them, but treat theories with varying levels of empirical support as roughly equivalent inputs. Their framework lacks a principled account of how much extrapolation a theory's current validation warrants. We supply this through a Bayesian account: theories earn extrapolation rights in proportion to their predictive success on humans, weighted by the surprise value of confirmed predictions. Bayne et al. (Bayne, Seth et al. 2024) provide a detailed taxonomy of consciousness tests and identify the circularity problem — validating theories requires tests, while validating tests requires theories — but do not connect this to the philosophy of science. Our Quinean grounding shows this circularity is a general feature of empirical science, resolvable through bootstrapping from paradigm cases via novel predictions. More fundamentally, neither account provides a criterion for distinguishing productive from idle questions about consciousness. Peirce's pragmatic maxim supplies this: questions whose answers would make no observable difference should be set aside not

Accepted at the AAAI 2026 Spring Symposium on "Machine Consciousness: Integrating Theory, Technology, and Philosophy."
© 2026 Association for the Advancement of Artificial Intelligence.

because they are unimportant, but because they are currently empirically inert.

The Grounding Problem

Consciousness as Family Resemblance

When researchers debate whether AI systems are conscious, they rarely specify what they mean by "consciousness." This is a problem as the term genuinely admits multiple distinct meanings (Bayne, Seth et al. 2024). Following Wittgenstein (Wittgenstein 1953), we should recognize that "consciousness" is a family-resemblance concept: there is no single essence that all uses share, but rather the meaning is composed of a network of overlapping similarities.

Consider the diverse phenomena that "consciousness" can denote:

- **Arousal/wakefulness**: The difference between deep sleep, wakefulness, and coma
- **Phenomenal quality**: The "what it's like" of experience (the redness of red, the painfulness of pain)
- **Unified experience**: How disparate sensory inputs combine into a single coherent scene
- **Access consciousness**: Information being available for verbal report and flexible behavioral control
- **Metacognition**: Thinking about one's own mental states
- **Self-modeling**: The sense of being an "I" having experiences
- **Valence**: The pleasure-pain dimension of experience

These are not merely different words for the same thing. They can dissociate. Patients with blindsight have visual information available for behavior (a form of access) without phenomenal experience of seeing (Weiskrantz 1986). Meditation practitioners report unified phenomenal experience with diminished self-modeling (Millière 2017). Anesthesia disrupts arousal while leaving questions about phenomenal experience contested.

So when someone asks "Is GPT-5 conscious?", which of these meanings do they intend? An LLM might have something analogous to access consciousness (information available for flexible use) while lacking anything resembling phenomenal quality or valence. Whether this counts as "conscious" depends on which meaning one privileges.

History of science suggests that to make progress, we need to isolate phenomena of interest and then subject it to theorization and experimentation. Hence, to make progress on the question of "consciousness", researchers first need to precisely identify what aspect of it are they targeting and only then make claims about it.

Of these aspects, we argue that phenomenal consciousness (or qualia) should be the priority target. Other members of the family (access consciousness, metacognition, attention) already have relatively well-understood functional decompositions and active research programs. Phenomenal consciousness is both the most resistant to functional explanation and arguably the most fundamental: even animals with minimal cortical capacity appear to possess phenomenal experience despite lacking access consciousness or self-models. It is also the aspect most directly relevant to moral consideration — a system with phenomenal experience may also have phenomenal valence (the felt quality of pleasure and pain), and if so, it can suffer. Targeting the hardest and most consequential member of the family first is the pragmatically correct strategy, precisely because it is where current approaches offer the least guidance.

The Problem of Idle Questions

Pragmatist philosophy offers a diagnostic criterion: **a question is idle if no possible observation could confirm or disconfirm the answer** (Peirce 1878/1992; James 1907). The classical "Does God exist outside spacetime and leave no detectable trace?" is idle in this sense. Nothing could count as evidence either way.

Many formulations of the AI consciousness question risk similar idleness. "Does this system *really* have phenomenal experience?" presupposes we have a way to detect phenomenal experience beyond behavioral and functional signatures. But if phenomenal consciousness is precisely that which cannot be observed from the outside, the question may lack empirical content.

This does not mean we should dismiss consciousness questions entirely. Brushing aside metaphysical puzzles don't make them go away and we risk ceasing a potentially productive line of inquiry early on. However, our methodology is compatible with multiple metaphysical interpretations. Realists can view it as tracking genuine phenomenal states. Illusionists can view it as modeling the mechanisms that generate consciousness-talk. We take no position on this dispute as the methodology is productive either way. Instead of asking whether an AI "really" has phenomenal experience, we can ask: Does it satisfy the functional criteria associated with consciousness in humans? Does it exhibit the neural (or computational) signatures that correlate with reported experience? These reformulated questions are empirically tractable[1].

The Need for Grounding

But a skeptic may argue that what is being studied is merely function and not consciousness. To counter this legitimate skepticism, we need to first ground our theories of consciousness on human experience. For humans, the primary observables are: (1) verbal reports of experience, (2) behavioral signatures like attention, discrimination, and flexible response, (3) neural correlates like EEG patterns, brain activity, and connectivity measures, and (4) responses to interventions like anesthesia, psychedelics, and brain stimulation. A theory of consciousness proves useful when it predicts systematic relationships among these observables.

Once we have isolated the aspect of consciousness we're interested in modeling, we proceed to make testable predictions from it. As those predictions get validated by human subjects, our confidence in them increases. If they fail, we go back to drawing board to iterate. This process of

[1] As a parallel consider the metaphysically loaded question: do atoms exist? When Greeks first posed it, atoms were mere speculation. What eventually resolved the question were accumulating evidence indicating that atoms indeed exist

conjecturing-testing-iterating is how all sciences progress, and the science of consciousness should not be any different.

Just as we need to have an empirically tested theory of fundamental particles before we could make predictions about what new kind of particles or properties exist, predictions of "consciousness" in novel configurations (such as our AIs) require grounding in specific, measurable phenomena. Crucially, this grounding must happen first for humans before extending to AI. We return to this point after examining why human observation is foundational.

Observation Requires Human Agreement

Quine on Observation Sentences

W.V.O. Quine's analysis of observation sentences illuminates why human consciousness must be the starting point for any science of consciousness, including machine consciousness (Quine 1960, 1969).

For Quine, observation sentences are "the starting point for our acquisition of knowledge" (Hylton and Kemp 2020). They are the sentences on which speakers of a language reliably agree subject to the same (experimental) context. "It's raining" is an observation sentence because competent speakers exposed to rain will assent to it.

Quine's key insight is that **all empirical evidence, no matter how sophisticated, ultimately connects to theory through observation sentences that depend on human perceptual agreement**. This isn't a normative statement of what constitutes evidence, but a Bayesian perspective on what ought to count as evidence. When multiple people agree on something, we eliminate our personal bias from observations and hence increase our confidence in a theory's general applicability. When a physicist reports that a detector registered a particle, this report is grounded in human observation of instrument readings and when the experiment replicates, our confidence in observed data goes up.

Applying This to Consciousness

For consciousness science, the observational base includes:

- Verbal reports ("I see red," "That was painful")
- Behavioral measures (discrimination performance, response times)
- Instrument readings (EEG traces, fMRI activations)
- Experimenters' observations of all the above

Notice that verbal reports occupy a privileged position. When a subject reports "I am experiencing a vivid mental image," this report is itself an observable. Other humans can hear or read it and agree on what was said. Of course, reports are not infallible; introspection can err (Schwitzgebel 2008) but we can design careful experiments that reduce noise/errors and via multiple repetitions across different subjects, we can statistically reduce the individual and isolate general properties of the underlying phenomena[2].

[2]Particle physicists routinely rely on and report statistical measures (like sigma deviation from mean); consciousness sciences can do the same.

For any scientific theory, empirical reports supply the requisite data to ground theory's predictions. Hence a science of consciousness must take human reports and observations as data, even while treating them as noisy measurements of underlying states. **Because our intuitions for what reliably counts as "consciousness" come from self-introspection and reports by other humans, we can't directly measure it for AI systems**.

Science proceeds when people agree to adopt a theory and hence an LLM's text outputs are not verbal reports in the same sense as a human. Barring philosophy departments, nobody doubts that other humans have internal experiences similar to their own. So, when a human reports "I feel pain," the level of belief that it actually feels like something from the inside is much higher than when an LLM reports that. To reiterate, we cannot simply assume that when an LLM outputs "I feel curious about this problem," it is reporting a genuine state analogous to human curiosity. The question of whether such outputs constitute genuine reports is precisely what we are trying to determine.

The Human Benchmark

Some argue we should apply a "Turing test" approach: if an AI's outputs are indistinguishable from those of a conscious human, we should attribute consciousness (Turing 1950). But this conflates behavioral equivalence with phenomenal equivalence. A system could produce human-like outputs through entirely different processes, just as a calculator and a human might produce the same arithmetic answers via different mechanisms.

More fundamentally, behavioral tests cannot establish consciousness because consciousness is not defined by behavior alone. A human could be sitting entirely still and yet daydreaming the most imaginative world from the inside. Similarly, a human in deep sleep or a coma would show neural activity on fMRI. The Chinese Room argument, whatever its other merits, correctly identifies that behavioral sophistication alone does not entail understanding or experience (Searle 1980).

This leaves us in an epistemically asymmetric position. For humans, we have:

- First-person access to our own experiences (though fallible)
- Reports from other humans whom we have reason to trust
- Evolutionary continuity suggesting similar mechanisms
- Neural correlates we can measure and manipulate

For AI systems, we have none of these. This asymmetry is not a temporary gap to be filled by better technology. It reflects the fundamental structure of evidence about consciousness that what we call as science requires.

The Human-First Methodology

Given the grounding problem and the primacy of human observation, we propose a methodological principle: *claims about AI consciousness should be derived from models first validated on human consciousness*.

The Core Methodology

The approach proceeds in five steps:

Step 1: Identify a specific, measurable phenomenon associated with consciousness in humans. Rather than tackling "consciousness" wholesale, target a tractable aspect: the neural correlates of visual awareness, the behavioral signatures of metacognitive access, the physiological markers of emotional valence, and so on.

Step 2: Build a predictive model of this phenomenon. The model should specify: given these inputs (stimuli, neural states, contexts), the system will produce these outputs (reports, behaviors, physiological responses). The model need not explain *why* consciousness exists. That is a separate, harder question (Chalmers 1995). But it should predict *when* and *how* specific conscious phenomena manifest.

Step 3: Validate the model empirically on humans. Test the model's predictions against human data[3]. Does it correctly predict when subjects report awareness? Does it account for the effects of attention, anesthesia, brain lesions? A model earns credibility through predictive success, not through intuitive appeal or theoretical elegance alone.

We do not propose a sharp threshold for 'sufficient validation.' Validation is better understood as a continuous Bayesian credence update: our confidence in a theory increases as its predictions are confirmed, weighted by how surprising those predictions are. A theory that merely retrodicts known phenomena — for instance, that attention modulates awareness or that anesthesia disrupts consciousness — earns modest credence, because these outcomes are expected under most theories. A theory that predicts genuinely novel effects earns substantially more. Consider an analogy: Eddington's 1919 confirmation of general relativity's prediction of light bending was paradigm-establishing not because light bending was important per se, but because the prediction was quantitatively precise and highly surprising under the Newtonian alternative.

Consciousness science awaits an analogous moment. Suppose a theory predicted that transcranial stimulation at a specific frequency, applied to a specific cortical region during a specific task, would cause subjects to report color inversion (experiencing red where green was presented). The inverted qualia scenario has long been treated as an idle philosophical thought experiment. A theory that renders it empirically tractable, and whose prediction is confirmed, would warrant a large credence update. Such theories should earn extrapolation rights: not by crossing a binary threshold, but by accumulating surprising predictive successes that tip collective scientific confidence.

Step 4: Apply the validated model to AI systems. As our confidence increases in a model of human consciousness, we can start applying it to AI. The model may then predict that certain AI architectures should (or should not) exhibit the relevant properties. Depending on our relative and absolute credence in different models of human consciousness, perhaps we will conclude that "consciousness" is substrate dependent or relies on quantum properties and hence current AI systems likely don't possess an internal experience. Or perhaps we will conclude that "consciousness" happens whenever there are feedback systems and hence we will be persuaded to conclude even a thermostat has an internal experience.

Step 5: Probe surprising predictions of model's extrapolation to AI systems. It's true that we have no way to directly verify any of the predictions about AI consciousness as a skeptic may dismiss any behavioral signature as "mere" functionality. However, this issue cuts right through the heart of the philosophy of science by asking how do we believe in things that we can't observe directly?

Our confidence in certain beliefs increase as we gather more data about different phenomena they are connected to (Quine 1969). Consider how physicists study black holes, entities that we can never observe directly. We're confident in existence of black holes because a highly validated theory (general relativity) postulates their existence *and* we also directly observed numerous surprising phenomena that black holes are related to (like accretion disks). Similarly, our confidence in attributing consciousness to an AI system should increase if our best theory of consciousness suggests it may have it *and* we also observe some surprising downstream or "emergent" behavioral phenomena that the theory says we should find.[4].

Why This Order Matters

One might object: why not develop models directly on AI systems? As we argued above, the answer lies in validation. For humans, we have multiple converging lines of evidence: reports, behavior, neural data, intervention effects. We can check whether a model's predictions about experience correlate with what subjects actually report. For AI, we have only behavior (outputs), and the question of whether these outputs reflect genuine experience is precisely what we are trying to determine.

Consider an analogy. Suppose we want to know whether a distant planet has water. We cannot directly sample the planet, but we can study spectroscopy on Earth, validate that certain spectral signatures indicate water, and then look for those signatures in the planet's light. The Earth-based validation is essential. Without it, we are just speculating about what the signatures mean.

Similarly, **validating consciousness models on humans provides the calibration needed to interpret AI systems**. If a model predicts that information integration above threshold X correlates with reported awareness in humans, we have reason to think that AI systems achieving similar integration might have analogous properties.

[3]We recognize the ethical issues with this, particularly for brain invasive methods. Hence, in the near term, non-invasive experiments are recommended but as brain-interfaces advance and become commonplace, the pace of experiments would accelerate

[4]It is hard to state upfront what these surprising signatures of AI consciousness could look like as our theories guide where to look. Nobody was looking for black holes before general relativity predicted them but once we had a theory, we could search for what it implied that wasn't already known or observed.

Implications for Current AI Consciousness Claims

Applying this methodology reveals that **current claims about AI consciousness are premature**. We do not yet have models of human consciousness validated well enough to license confident extrapolation to AI (Seth and Bayne 2022). Integrated Information Theory and Global Workspace Theory are steps in the right direction, but we're still quite early when it comes to their empirical support (Tononi et al. 2016; Baars 2005; Dehaene and Changeux 2011).

This does not mean AI consciousness is impossible or that research should stop. Rather, it suggests redirecting effort toward:

- Developing and testing human consciousness models with greater rigor
- Building AI systems whose architectures parallel human consciousness-related mechanisms (for comparative study)
- Designing experiments that can discriminate between theories

Our goal should be to earn the right to make claims about AI consciousness by first understanding consciousness where we can actually measure it.

Addressing Objections

Objection 1: The methodology is circular: validating models require knowing what consciousness is, which is what the models are supposed to tell us

This circularity is benign and unavoidable in all science. We bootstrap by starting with paradigm cases on which there is broad agreement (normal waking human experience is conscious; dreamless sleep is not; blindsight dissociates access from phenomenal report). Models are initially validated against these clear cases, then extended to contested ones. The circle is not vicious because it spirals outward: successful predictions in clear cases license application to unclear cases, generating new data that refines the model. This is how all science proceeds: from agreed-upon observations to theoretical extension.

Objection 2: If a consciousness theory is validated on humans, do we accept its wild implications such as a network of XOR gates being conscious?

If our best-validated theory implies that certain simple systems have minimal conscious experience, we should take this seriously[5]. The history of science shows that our pretheoretic intuitions about what is possible are often wrong. Our methodology provides principled grounds for revising intuitions: we accept surprising implications when they follow from theories with strong predictive track records, and reject them when theories fail empirical tests. The alternative—rejecting theories because their implications seem strange—would have blocked most scientific progress.

Objection 3: Waiting to validate theories on humans could delay recognizing genuinely conscious AI.

We acknowledge this concern and propose a graded precautionary framework. Rather than treating moral caution as binary — either assume consciousness or don't — we recommend calibrating precaution to the number of validated indicators a system satisfies (Butlin, Long et al. 2023; Bayne, Seth et al. 2024). A system satisfying zero indicators derived from human-validated models (e.g., a simple calculator) warrants no special moral consideration. A system satisfying many indicators (e.g., a hypothetical AI with recurrent processing, global workspace dynamics, and metacognitive self-reports) warrants substantial caution.

We note that the asymmetry of moral risk favors caution: the cost of under-attribution (ignoring genuine suffering) is plausibly higher than the cost of over-attribution (extending unnecessary moral consideration). Hence, where indicator evidence is ambiguous, we recommend erring toward moral consideration rather than neutrality.

Objection 4: What if multiple divergent models successfully predict the same aspect of consciousness?

Resolution of competing theories happens over time, so we don't need to immediately resolve which one is correct. Each theory is a "bet" and over time such bets cash out via their successful, novel predictions. As an example, the heliocentric and epicycles theory of planetary motion co-existed and competed until one of them won out by virtue of accumulative evidence.

Objection 5: Why not start with animals whose consciousness is likely to be "simpler" and hence easier to model/understand?

Our recommendation to ground initial theories of consciousness on humans is motivated purely from the pragmatic view that it's harder to deny internal experiences among fellow humans than for animals. Moreover, the verbal reports via humans provides richer set of data which would be missing for purely behavioral data from animals. In fact, the human-first methodology also addresses skepticism about animal consciousness. Once we have validated models of human consciousness, we can apply them to non-human animals with similar neural architectures and provide principled grounds for attributing consciousness to creatures who cannot verbally report.

Objection 6: Human consciousness might be substrate-specific; validating on humans tells us nothing about AI.

This concern (Tegmark 2017) admits three distinct versions with different implications.

- *Weak substrate specificity*: biological systems have specific functional properties (continuous dynamics, metabolic self-maintenance, particular processing timescales) that happen to be necessary for consciousness, but could in principle be replicated in other substrates. Our methodology handles this directly: validated models would reveal which functional properties matter, and what sort of machines could possess them.
- *Moderate substrate specificity*: specific physical properties of biological systems (e.g., quantum coherence, molecular-scale dynamics) are necessary for consciousness but produce detectable functional downstream effects. Our methodology handles this: we would observe that certain predictions systematically fail when applied

[5]Just as physicists took seriously the counterintuitive implications of quantum mechanics and relativity.

to AI, pointing toward missing physical properties.

- *Strong substrate specificity* (the zombie scenario): consciousness requires biology, with zero detectable functional differences between conscious and zombie systems. Here our methodology cannot help, but also no methodology can. If there are truly no observable differences between conscious and non-conscious systems, then by the pragmatic maxim, the question has become empirically idle.

Importantly, our methodology remains maximally informative across all three scenarios: in the weak and moderate cases, it helps discover what matters; in the strong case, it reveals when we have reached the epistemic wall.

Objection 7: Reports are unreliable; we cannot trust human verbal reports about experience.

Reports are indeed imperfect (Schwitzgebel 2008). But science routinely works with noisy data. The solution is not to abandon reports but to model them as noisy measurements of underlying states, triangulating across multiple sources: reports, behavior, neural data, intervention effects. This is standard methodology in psychophysics and cognitive neuroscience.

Conclusion

The question "Is this AI conscious?" cannot be answered, indeed cannot even be made precise, without validated models of consciousness grounded in human data. Current debates about AI consciousness are largely unproductive because they lack this grounding: participants use "consciousness" in different senses, appeal to untested theories, and have no empirical basis for resolving disagreements.

We have argued for a human-first methodology: identify specific consciousness-related phenomena in humans, build predictive models, validate them empirically, and only then apply them to AI. This approach transforms idle philosophical debate into productive science. It respects the epistemic asymmetry between human and machine consciousness while providing a path toward eventual resolution.

The field of AI consciousness is pre-paradigmatic right now. Like atomic theory before Dalton or germ theory before Koch, it awaits the validated models that will make it a genuine science. The paradigm-establishing moment will not be a single dramatic experiment, but a pattern of predictive success across domains: a theory that simultaneously explains why certain neural signatures correlate with reported awareness, why specific interventions abolish consciousness, why certain disorders produce specific experiential deficits and that makes a genuinely novel prediction about an unexplored condition that proves correct. A theory that could predict, for instance, the precise conditions under which a subject would report color inversion — rendering a long-standing philosophical thought experiment empirically tractable — would exemplify the kind of surprising success that tips collective confidence, much as Eddington's confirmation of light bending tipped physics toward general relativity. Building toward that moment requires first understanding consciousness where we can actually observe it: in ourselves.

Acknowledgments

We thank Dhruv, Diksha and the anonymous reviewer for their comments and feedback on the paper.

References

Baars, B. J. 2005. Global Workspace Theory of Consciousness: Toward a Cognitive Neuroscience of Human Experience. *Progress in Brain Research*, 150: 45–53.

Bayne, T.; Seth, A. K.; et al. 2024. Tests for Consciousness in Humans and Beyond. *Trends in Cognitive Sciences*, 28(5): 454–466.

Butlin, P.; Long, R.; et al. 2023. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708.

Chalmers, D. J. 1995. Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies*, 2(3): 200–219.

Dehaene, S.; and Changeux, J.-P. 2011. Experimental and Theoretical Approaches to Conscious Processing. *Neuron*, 70(2): 200–227.

Dennett, D. C. 1996. Facing Backwards on the Problem of Consciousness. *Journal of Consciousness Studies*, 3(1): 4–6.

Dewey, J. 1929. *The Quest for Certainty*. Minton, Balch & Company.

Hylton, P.; and Kemp, G. 2020. Willard Van Orman Quine. In *Stanford Encyclopedia of Philosophy*.

James, W. 1907. *Pragmatism: A New Name for Some Old Ways of Thinking*. Longmans, Green.

Millière, R. 2017. Looking for the Self: Phenomenology, Neurophysiology and Philosophical Significance of Drug-Induced Ego Dissolution. *Frontiers in Human Neuroscience*, 11: 245.

Peirce, C. S. 1878/1992. How to Make Our Ideas Clear. In *The Essential Peirce*, volume 1. Indiana University Press.

Quine, W. V. O. 1960. *Word and Object*. MIT Press.

Quine, W. V. O. 1969. *Ontological Relativity and Other Essays*. Columbia University Press.

Schwitzgebel, E. 2008. The Unreliability of Naive Introspection. *Philosophical Review*, 117(2): 245–273.

Schwitzgebel, E. 2025. AI and Consciousness. arXiv:2510.09858.

Searle, J. R. 1980. Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3(3): 417–424.

Seth, A. K. 2016. The Hard Problem of Consciousness is a Distraction from the Real One. https://aeon.co/essays/the-hard-problem-of-consciousness-is-a-distraction-from-the-real-one. Aeon, November 2016.

Seth, A. K.; and Bayne, T. 2022. Theories of Consciousness. *Nature Reviews Neuroscience*, 23(7): 439–452.

Tegmark, M. 2017. *Life 3.0: Being Human in the Age of Artificial Intelligence*. Knopf.

Tononi, G.; Boly, M.; Massimini, M.; and Koch, C. 2016. Integrated Information Theory: From Consciousness to Its Physical Substrate. *Nature Reviews Neuroscience*, 17(7): 450–461.

Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind*, 59(236): 433–460.

Weiskrantz, L. 1986. *Blindsight: A Case Study and Implications*. Oxford University Press.

Wittgenstein, L. 1953. *Philosophical Investigations*. Blackwell.