The Mimicry Trap

How We Define Intelligence to Exclude Inconvenient Minds

Giorgio F. Gilestro, PhD

Associate Professor in Systems Neurobiology
Department of Life Sciences, Imperial College London, UK

giorgio@gilest.ro · g.gilestro@imperial.ac.uk

Abstract

Short abstract. When entities presumed incapable of intelligence produce qualifying behavior, a recurrent move concedes the performance but denies its evidential force, redescribing it as imitation, simulation, or surface mimicry. I call this the *mimicry trap* and trace it across three otherwise disanalogous domains: human reception (the eighteenth-century reader Phillis Wheatley), animal cognition, and current debate over Large Language Models. The framework names five diagnostic markers—failure to specify falsifiers, asymmetric standards, shifting criteria, mechanism-based dismissal, and appeals to a missing essence. Applied to LLMs and read against Ockham's razor, blanket denial loses its claim to be the cautious default.

Long abstract. When entities presumed incapable of intelligence produce behavior that would ordinarily count as evidence of it, a recurrent response is to concede the performance while denying its evidential force: the behavior is redescribed as imitation, simulation, contamination, or surface patterning. I call this structure the *mimicry trap*. The article argues that the trap is a cross-domain epistemic pattern that recurs wherever intelligence is at stake in an unfamiliar substrate: in humans of disfavored groups (the eighteenth-century reception of Phillis Wheatley), in non-human animals (anthropodenial and anthropofabulation in comparative cognition), and in machines (contemporary debates over Large Language Models). Existing labels—*anthropodenial*, *anthropofabulation*, and the *AI Effect*—capture local instances; the present framework identifies their shared diagnostics: failure to specify falsifiers, asymmetric standards, shifting criteria, mechanism-based dismissal, and appeals to an unobservable missing essence. Applied to LLMs, the framework distinguishes legitimate skepticism about grounding, agency, embodiment, and benchmark validity from blanket denial that cannot say what evidence would change its verdict. I argue that recent behavioral and mechanistic evidence—including recoverable world models, algorithmic circuits, calibrated self-assessment, and high performance on some theory-of-mind and mathematical-reasoning tasks—raises the evidential cost of crude "mere mimicry" accounts. A Bayesian reconstruction of Turing's imitation game treats behavioral performance and mechanistic evidence as jointly relevant to intelligence attribution. Finally, drawing on Ockham's razor, I argue that "real understanding," "true semantic content," and "genuine intelligence" are legitimate theoretical posits only if they make independent empirical or explanatory differences; otherwise they function as non-identifiable entities preserving a prior verdict.

Keywords: anthropodenial, artificial intelligence, functionalism, intelligence attribution, large language models, mimicry trap, Ockham's razor, philosophy of mind

1 Introduction: The Recurrence of the Mimicry Argument

Several influential lines of contemporary AI skepticism share a recurring methodological structure—one that is older than artificial intelligence. When Thomas Jefferson encountered the poems of the enslaved teenager Phillis Wheatley in 1773, he did not dispute that the poems were technically accomplished; he denied that they could be evidence of intellect, on the explicit ground that intellect was not the kind of property that could belong to the kind of person Wheatley was. The same move was named "anthropodenial" by the primatologist Frans de Waal at the close of the twentieth century [de Waal, 1999]: the *a priori* rejection of hypotheses ascribing human-like cognitive capacities to non-human animals, regardless of the behavioural evidence. Cameron Buckner identified its complementary inflation, "anthropofabulation" [Buckner, 2013]: the elevation of human performance to set unrealistic bars for non-human candidates. Pamela McCorduck [McCorduck, 2004] and Larry Tesler informally codified the early artificial-intelligence version as the "AI Effect" or Tesler's Theorem—"AI is whatever hasn't been done yet"—and Mitchell and Krakauer [Mitchell & Krakauer, 2023] have catalogued the contemporary LLM form as "moving goalposts." The same structure is visible today in responses to GPT-4 passing the Uniform Bar Examination [Katz et al., 2024] and to AlphaGeometry solving International Mathematical Olympiad problems [Trinh et al., 2024]: not, in any of these cases, denial of the performance, but denial that the performance constitutes evidence of the capacity. Each of these accounts captures a piece of the same argumentative structure operating in a different domain.

The denial itself can land at two distinct places. Sometimes it lands on the evidence: the benchmark, on inspection, does not measure the capacity, or measures it imperfectly—an objection that is contestable on its merits and in principle responsive to a better test. Sometimes it lands earlier, at the level of what would count as evidence in the first place: statistical learning, however successful, is held not to be the kind of process from which reasoning can in principle emerge; intellect is not the kind of property that could belong to the kind of person Wheatley was; tool use is not the kind of behaviour that could indicate the kind of cognition we attribute to children—denials that, by their own structure, no behavioural evidence could move. The first kind operates at the level of evidence and remains open to revision; the second operates at the level of priors and is, by construction, unfalsifiable.

This paper takes a broadly functionalist working definition of intelligence—the capacity to achieve goals across a wide range of environments through flexible, appropriate behaviour (§2)—and reads Turing's 1950 imitation game as the canonical attempt to discipline this kind of dispute by tying intelligence-attribution to publicly assessable behaviour and refusing to let prior-level commitments enter through the back door. The test was mechanism-blind by design—in 1950 the only honest way to operationalise the question—but mechanism has since become inspectable through modern interpretability methods. The paper's central technical move is a Bayesian reconstruction of the imitation game (§4) that updates Turing's framework accordingly: behavioural performance and mechanistic evidence update intelligence-attribution jointly and symmetrically, while the prior is forced into the calculation explicitly rather than smuggled in case by case. The reconstruction makes the two-step distinction of the previous paragraph crisp: evidence-level objections move the posterior and remain in the empirical arena; prior-level objections refuse to update at all, and when no observation could in principle move them, they have left the empirical arena entirely.

The pattern that unites these cases—conceding performance while denying its evidential significance on grounds that derive from what the candidate *is* rather than from anything the candidate did—this article calls *the mimicry trap*. Naming the pattern does substantive work. Once it is in view, specific intelligence-denial arguments can be tested against the diagnostic developed in §2.2: a seven-marker checklist whose *structural* tier (falsifiability, invisible absence,

ontological precedence) is on its own diagnostic, and whose *symptomatic* tier (consistency, goalpost, mechanism, contamination) describes patterns that become diagnostic when deployed wholesale. The diagnostic distinguishes *disciplined* skepticism—which specifies in advance what evidence would change its verdict—from positions that, by their own structure, cannot; and applied case by case in Table 2 (§9), it consolidates the historical and contemporary instances into a single comparative view.

The diagnostic has methodological consequences. The default shifts with it: blanket denial of LLM intelligence, and the undisciplined agnosticism that refuses to specify what evidence would update it, are no longer the cautious starting point once the cumulative empirical record is in view; *disciplined* agnosticism about specific richer notions (grounded semantics, autonomous agency, consciousness) remains legitimate. Read against Ockham's razor in its classical form—*entia non sunt multiplicanda*—and Lindley's Cromwell rule [Lindley, 1985], the residual skeptical work turns out to be done by entities that are non-identifiable in the Bayesian sense: a structural rather than rhetorical result.

2 Defining Intelligence Functionally

This paper restricts its analysis to questions of *intelligence* and does not make claims about phenomenal consciousness, subjective experience, or moral status. The restriction is methodological, not dismissive. Intelligence is in practice the first proxy by which consciousness is reasoned about: the entities to which we are tempted to ascribe consciousness are typically entities to which we have already ascribed intelligence, and the difficulty of recognising intelligence in an unfamiliar substrate is, structurally, the same difficulty that recurs when the discussion moves to consciousness. The framework developed here could in principle be extended to that adjacent debate once the structural moves it catalogues are recognised in operation there, but the extension is not attempted in this paper; its task is to discipline the first step. One can in any case coherently hold that LLMs exhibit functional intelligence while remaining agnostic about whether they have subjective experiences, just as one might attribute intelligence to a corporation or a distributed system without attributing consciousness to it [Schwitzgebel & Garza, 2015].

I adopt a broadly functionalist working definition of intelligence: the capacity to achieve goals across a wide range of environments, particularly novel ones, through flexible, appropriate behavior (Legg & Hutter, 2007; Chollet, 2019). The definition has three virtues. It is *substrate-neutral*: it does not presuppose biological neurons or any particular physical medium. This is not a stylistic preference but a structural one: functionalism is what remains of intelligence once the parts of any candidate definition that are axiomatic, tautological, or substrate-biased—that is, the parts that cannot be falsified by anything a system does or could do—are stripped away. It is *operationalisable*: tests can in principle assess flexible goal-achievement across novel environments, even granting Chollet's (2019) point that most current benchmarks measure skill acquisition rather than generalisation. And it is *continuous with comparative practice*: it aligns with how we actually attribute intelligence across the biological world. We count octopuses intelligent because they solve novel problems and exhibit flexible behavior [Godfrey-Smith, 2016], and corvids on the basis of causal reasoning performance [Taylor, 2014], not from any verified "intelligence substance." We must nonetheless acknowledge a deeper limit: our concept of intelligence is anthropo-limited. We recognize the octopus partly because its tasks (opening jars, navigating mazes) map onto human capabilities; fish, by such metrics, look unintelligent, even though their three-dimensional spatial memory and electromagnetic sensitivity may be genuine cognitive capacities our evaluation frameworks systematically miss. If we struggle to recognize intelligence that evolution has produced in alien substrates over hundreds of millions of years, we should be cautious about confident pronouncements regarding intelligence in substrates we

have only recently created.

The paper does *not* claim that all legitimate uses of "intelligence" must be functionalist, that functionalism is the correct general theory of mind, or that no richer concepts of cognition do any work. It claims that in the contexts under discussion (comparative cognition, AI capability assessment, behavioral attribution across substrates), functional intelligence is the operative notion, and that critics who deny intelligence to functionally satisfying systems owe an alternative concept that is testable. The requirement is non-negotiable: a rival definition must specify what additional property is required, why it belongs to *intelligence* rather than to consciousness, agency, embodiment, normativity, or moral status, and what observable evidence could in principle establish or refute its presence in a candidate system. A concept whose application conditions cannot themselves be specified, or whose verdict no conceivable observation could revise, is not yet a candidate definition of intelligence; it is the prior commitment that the diagnostic of §2.2 will register. This is not a stipulation that functionalism wins; it is a procedural requirement on what any rival must deliver before it can compete.

Two further scope-clarifications matter. First, "intelligence," "understanding," "semantic content," "cognition," "theory of mind," "meta-cognition," and "agency" are related but not identical concepts; a skeptic may reasonably accept functional intelligence while denying understanding or grounded semantics, and the paper takes *functional intelligence* as its target while treating the others as contested neighbours. Where empirical work on theory-of-mind operationalisations is reviewed (§5.5), the strict claim is high performance on those operationalisations, with the question of whether high performance constitutes possession of theory of mind left undecided in the steelman of §5.6. Consciousness in particular has recently been the subject of a BBS target article advancing biological naturalism [Seth, 2025]; the present paper takes no view on that thesis, which falls on the consciousness side of the scope distinction already drawn. Second, the functionalist definition requires generality *across* environments rather than competence within one, and natural-language environments are themselves heterogeneous (mathematics, code, conversation, planning, fiction, scientific argument)—the cross-domain transfer documented in §5 is empirically across environments, though the steelman of §5.6 grants the contestability of that reading.

2.1 Map of Claims

With the target concept of functional intelligence in place, the argument that follows is best read as a hierarchy of nested claims, of decreasing generality and increasing contestability.

1. *Minimal claim* [§2.2]. Some intelligence-denial arguments, as deployed in the contemporary AI debate, are unfalsifiable because they cannot specify in advance any evidence that would change their verdict.

2. *Cross-domain claim* [§§3–4]. The same argumentative structure recurs across cases that are otherwise wholly disanalogous: anthropodenial in animal cognition, the eighteenth-century reception of Phillis Wheatley, and contemporary skepticism about LLMs. The structure is what this paper calls the mimicry trap.

3. *LLM-specific claim* [§4]. Several influential mimicry-skeptical arguments about LLMs (the "stochastic parrots" framing, the next-token reductio, blanket data-contamination dismissal, Floridi's "agency without intelligence" thesis) instantiate this structure.

4. *Evidential claim* [§5]. The cumulative behavioral and mechanistic record (world models, algorithmic circuits, theory-of-mind performance, calibrated meta-cognitive prediction) is incompatible with crude mimicry accounts of LLM behavior, even granting that it does

not, by itself, settle every harder question about understanding, semantics, agency, or consciousness.

5. *Burden-shift claim* [§7]. Given (3) and (4), blanket denial of LLM intelligence now owes positive argument; the inherited default of absence is no longer doing its old work, even though disciplined agnosticism about specific richer notions remains in good standing.

The minimal and cross-domain claims are the paper's strongest. The LLM-specific and evidential claims are well-supported but more contestable. The burden-shift claim is the most contested and is qualified accordingly. A reader who accepts the minimal and cross-domain claims, but reserves judgment on (4) and (5), takes much of what the paper has to offer.

2.2 The Mimicry Trap: Structure and Diagnostic Framework

With the target concept in place, the framework can be put in front of the reader as an explicit instrument, so that the historical, comparative, and contemporary cases that follow can be read through it rather than re-deriving the diagnosis each time.

The mimicry trap, as this paper uses the term, is the conjunction of three elements: (i) a prior commitment, often substrate-based, that some entity cannot possess a given cognitive capacity; (ii) the appearance, in that entity's behavior, of evidence that would be taken as supporting the capacity if produced by another entity; and (iii) an interpretive procedure that reclassifies the inconvenient evidence (as imitation, simulation, contamination, or surface pattern-matching) so that the prior commitment is preserved. Each element on its own can be a legitimate move; the diagnostic question is whether (iii) is operating in such a way that no possible evidence could move the prior. When that is so, the position is no longer responsive to the world, and what looked like an empirical hypothesis turns out to be a stipulation in empirical clothing. Table 1 summarises the seven diagnostic tests deployed in what follows. Each test has been articulated in some form by prior work—falsifiability by Popper, with its Bayesian formalisation as Lindley's *Cromwell's rule* [Lindley, 1985]; consistency by Buckner [Buckner, 2013]; the goalpost pattern by McCorduck [McCorduck, 2004], Tesler under the heading of the AI Effect, and Mitchell & Krakauer [Mitchell & Krakauer, 2023]; the contamination pattern by de Waal [de Waal, 1999]—and the contribution offered here is consolidation: gathering criteria that exist separately into a single applicable instrument and sharpening the diagnosis of unfalsifiability and tautology developed in the body of the paper.

The seven tests are offered not as a closed taxonomy but as a diagnostic proposal: commentators may reject specific tests, refine the formulations, or extend the list, and the framework will have done useful work even where its particulars are revised. The two tiers carry different evidential weight. Failure of any one of the *structural* tests is on its own diagnostic, because each describes the trap form directly: an unfalsifiable position, an unobservable missing essence, or a conclusion fixed by what the entity *is* rather than by what it does. Within the structural tier, ontological precedence is the most general member—the form to which the others ultimately reduce, since unfalsifiability and invisible-absence appeals are species of conclusion-fixed-by-prior-commitment. Failure of a *symptomatic* test, by contrast, can be principled in any single instance — a single mechanism-based or contamination-based objection can be valid, a single benchmark revision can be reasonable updating — and the diagnosis applies only when symptomatic moves are deployed wholesale, regardless of whether any specific instance has been empirically discharged. The compressed form of the structural tier is the *concession test*, a single direct challenge any mimicry-skeptical position must be able to answer: *imagine a continuation of the trajectory of LLM capability that has held over the past five years—greater accuracy on novel tasks, more sophisticated internal representations under mechanistic probing, behavioral performance on*

Test	Question to Ask	Red Flag
Structural — single failure is diagnostic		
Falsifiability	What evidence would change your mind?	No test could ever demonstrate intelligence; criteria rejected once met
Invisible Absence	Is an unobservable quality claimed missing?	Outputs match intelligent agents, but "essence" denied
Ontological Precedence	Does the conclusion follow from what the entity *is*?	Argument guarantees conclusion regardless of evidence
Symptomatic — pattern required		
Consistency	Would you apply this standard to humans?	Different standards for AI vs. biological systems
Goal-Post	Have criteria shifted after being met?	Previously accepted benchmarks dismissed once passed
Mechanism	Is the objection about *how* rather than *what*?	Performance dismissed due to underlying process
Contamination	Is learning treated as disqualifying?	Training data disqualifies, but human education doesn't

Table 1: The diagnostic tests, organised by evidential weight. Structural tests describe the trap form directly (single failure diagnostic); symptomatic tests describe patterns that become diagnostic only when applied wholesale.

theory-of-mind, mathematical reasoning, calibrated self-prediction, and out-of-distribution generalisation matching or exceeding that of human experts. At what point would you concede that what is being witnessed is no longer mimicry? What would the system have to demonstrate, and what specific empirical signature would suffice? Three constraints discipline the answer. The criterion specified must be:

1. *operational* — expressible as a test that could in principle be run, with a measurable outcome (not "genuine understanding," "real semantic engagement," or any locution whose application conditions cannot themselves be specified);
2. *consistent* — the same criterion, if met by a human or an animal, would also count as evidence of intelligence in that case (not a bar deliberately set above human performance);
3. *specified in advance* — fixed before the evidence arrives, with the commitment that the verdict updates if the criterion is met (not retrofitted as a new objection after each previous criterion has been satisfied).

A reply that meets all three constraints has earned the right to its skepticism: it is doing science. A reply that fails any of them—that retreats into non-operational vocabulary, sets asymmetric bars, or constructs new criteria after old ones are met—is the case to which the diagnosis applies. The function of the framework is not to invalidate skepticism (the steelman of §5.6 treats it as substantive in several of its forms) but to separate skepticism that is responsive to evidence from skepticism that is not.

3 Historical and Comparative Precedents

The framework of §2 is contemporary in its terminology, but the structural pattern is older than artificial intelligence and recurs across cases that are otherwise wholly disanalogous. This section reviews two: anthropodenial in twentieth-century comparative cognition (§3.1), and the eighteenth-century reception of Phillis Wheatley's poetry (§3.2). Together they show that the trap is not a feature of any particular substrate-based prior but of a particular kind of unfalsifiable interpretive procedure.

3.1 Anthropodenial in Comparative Cognition

For decades, scientists systematically underestimated animal cognitive capacities, dismissing behavioral evidence of reasoning, emotion, and planning as anthropomorphic projection. De Waal (1999) coined "anthropodenial" for the *a priori* rejection of hypotheses ascribing human-like cognitive features to animals regardless of behavioral evidence: "the simplest, most parsimonious view is that if two related species act similarly under similar circumstances, they must be similarly motivated" [de Waal, 2016]; denying that motivational similarity in the face of behavioral similarity requires positing the kind of unobservable difference §7.1 will analyze as a multiplied *ens*. Buckner (2013) named the complementary bias "anthropofabulation" [Buckner, 2013]: setting unrealistic bars for animal (or artificial) cognition based on an inflated view of typical human performance, while in fact human cognition relies heavily on heuristics and pattern-matching that, accurately described, sound much like the "mere statistical correlation" attributed dismissively to AI. The two work in concert, inflating human capacities while deflating evidence of similar capacities elsewhere. Shevlin and Halina (2019) make the methodological consequence explicit: "rich psychological terms" such as understanding and theory of mind require careful and *consistent* application across both biological and artificial systems [Shevlin & Halina, 2019].

Two cases make the pattern recognisable. The trajectory of corvid cognition is the more familiar one: New Caledonian crows manufacture hooked tools, bend wire into hooks for novel retrieval problems, and execute multi-step sequences in which one tool retrieves another [Taylor, 2014]. Each demonstration was met with criteria adjustment (instinct, then conditioned reflex, then domain-specific adaptation, then meta-tool use), and corvids are now broadly recognized as causal reasoners only after decades of dismissal whose argumentative structure is exactly the one this paper diagnoses. The more striking case, which I approach with both relevant familiarity and a professional stake to declare since insect cognition is the subject of my own laboratory's research, concerns bees. Bee brains contain on the order of a million neurons, an architecture once considered too simple to support anything beyond reflex. Loukola et al. [Loukola et al., 2017] demonstrated that bumblebees could be trained, by demonstrator observation, to roll a ball into a target for sucrose reward, and crucially that observers did not slavishly imitate but generalized: tested without the demonstrator, they rolled the closest ball rather than the further one they had seen moved. Subsequent work has documented apparent play behavior, social transmission of learned techniques, and decision biases mirroring affective states observed in vertebrates [Chittka, 2022]. Each finding has been met by the same protective move—insects are reflexive, their cognition must be reducible to simple rules, any apparent flexibility is anthropomorphic projection—and the structural parallel to "LLMs are merely autocompletion" is direct and unforced.

The methodological point generalises directly to artificial systems: if behavioral evidence suffices to attribute intelligence to biological systems with alien neural architectures, it should count as evidence (not proof but evidence) when produced by silicon systems as well.

3.2 Phillis Wheatley

The second precedent is human and predates artificial intelligence by two centuries. Phillis Wheatley, an enslaved African woman in Boston, composed by her early teens sophisticated Neoclassical poetry that brought her transatlantic attention [Gates, 2003]. The work posed an ideological problem precisely because it was accomplished: the justification of African slavery rested substantially on claims of African intellectual inferiority, and an enslaved African woman producing poetry of European quality threatened that justification. In 1772 Boston printers refused to publish her not because the poetry was deficient but because they did not believe an

African woman could have written it. Wheatley was subjected to an oral examination before a committee of eighteen prominent Bostonians (including Governor Hutchinson and John Hancock), who interrogated her on her knowledge of the classics; they signed an "attestation" of her authorship, printed as a preface to her 1773 *Poems on Various Subjects, Religious and Moral*.

The examination's existence already reveals an asymmetric burden of proof: white poets were not routinely subjected to tribunals verifying their authorship. But the deeper problem is axiomatic. The denial of African intellectual capacity was not an empirical hypothesis Wheatley's performance could refute; it was an ontological commitment determining how any evidence would be interpreted. Within such a framework, any apparent demonstration of intelligence by an African must, by logical necessity, be apparent only. The eighteen signatories could attest that she had written the poems but not that she possessed genuine poetic intelligence, because that category had been defined to exclude her *a priori*. Thomas Jefferson's assessment in *Notes on the State of Virginia* (Query XIV, 1785) shows the unfalsifiable commitment in operation [Jefferson, 1785]. Jefferson did not argue the poetry was technically deficient; he could not, because it was not. Instead: "Religion, indeed, has produced a Phillis Wheatley; but it could not produce a poet. The compositions published under her name are below the dignity of criticism." The grammatical structure is the diagnostic feature: Wheatley is conceded to have been "produced"; what is denied is that she is a poet. *The poems exist; the poet does not.* What evidence could refute the position? Her poetry already displayed sophisticated classical allusion, precise metrical control, and thematic depth; if these are insufficient, what would suffice? Within Jefferson's framework, nothing. The deficiency he posits is not observable in the work; it is inferred from his prior commitment about what an African could and could not genuinely possess. The "absence" he detects is one he brought to the reading. A characteristic accompaniment is the treatment of environmental exposure as evidence *against* capacity rather than as its precondition: Wheatley's education and library access became reasons to deny the authenticity of her intelligence ("mere absorption"). A structurally similar move appears in blanket versions of the contemporary data-contamination objection (§5.4).

3.3 What These Cases Show

Across these otherwise disanalogous cases the same interpretive structure recurs: a prior commitment about who or what can possess a given capacity; performances from the favored class taken at face value while similar performances from the excluded class are discounted; and a protective redescription of the inconvenient evidence as "mere" imitation, instinct, conditioning, absorption, or mechanism.[1]

The crucial question—*does the performance exhibit the functional marks of the capacity?*—is displaced by another: *is this the kind of entity to which the capacity may be attributed?* Once the displacement occurs, the prior becomes effectively unfalsifiable. Wheatley's poetry is absorbed rather than written; corvid tool use is instinct rather than causal reasoning; bee social learning is reflexive rule-following rather than cognition. The performance remains visible, but its evidential force is neutralised.

The lesson is not that every surprising performance proves the contested capacity. It is that evidence must be allowed to count as evidence: skepticism is legitimate when it offers a better explanation of the performance, defective when it functions as a rule that no performance by the excluded subject could ever warrant the attribution. The contemporary version is what §4

[1]The Wheatley case is morally charged, and the analogy made here is purely structural. The wrongs done to enslaved Africans are categorically incommensurable with anything at issue in LLM debates; the parallel concerns the formal structure of an interpretive error, not the parties, stakes, or moral situations of the cases. The argument that contemporary mimicry-sceptics are wrong, where it is made, rests on the analysis of §4–§7.1, not on the historical analogy.

takes up.

4 The Contemporary Instantiation: From Turing to Next-Token Predictors

The historical cases of §3 share a methodological deficit: the absence of a publicly assessable functional criterion for intelligence-attribution. Turing's 1950 imitation game can be read as the canonical attempt to supply one—not to settle the metaphysics of thinking, but to discipline intelligence-attribution by tying it to behavioral performance. Pose the question that way—if the system's outputs are indistinguishable from those of an intelligent agent, that is evidence of intelligence, whatever the system happens to be made of—and the substrate-bound and ontological criteria that drove the historical errors lose their grip.

The test is complete within its declared scope. In 1950, the kinds of mechanism through which an artificial agent might exhibit intelligence were effectively unpredictable, and an explicitly mechanism-blind criterion was the only honest way to operationalize the question. The contemporary mimicry response goes the other way: rather than adding mechanism as a complementary channel of evidence to behavior, it uses construction history (the system is, after all, "only" a next-token predictor over text) to nullify the behavioral evidence Turing's test was designed to make count. That is a reversal of Turing's discipline, not a refinement of it.

A historical concession is in order before the reconstruction. Turing's §5 ("Arguments from Various Disabilities"), often read as a rhetorical clearing exercise, already contains a structural diagnosis of the family of moves this paper systematises. The catalog of capacities machines "will never" display ("be kind, resourceful, beautiful, friendly, have initiative ... do something really new") is the goal-post pattern set out in advance of its own falsification, and Turing's diagnosis—that such claims are "mostly founded on the principle of scientific induction" from the calculators of the day to all possible machines—is the asymmetric inductive standard later named *anthropodenial* [de Waal, 1999]. The mechanism-dismissal pattern is named explicitly: "the criticisms we are considering here are often disguised forms of the argument from consciousness ... the method (whatever it may be, for it must be mechanical) is really rather base." The cross-domain consolidation of §3 is sketched in a single passing aside placing the man–machine asymmetry on a continuum with racial othering. And the distinction Turing draws between "errors of functioning" (architectural artefacts of how the system is built) and "errors of conclusion" (substantive cognitive failures on the task being asked) anticipates a contemporary class of objections that conflate the two—the contention that an LLM cannot really reason because it miscounts the letters in "strawberry" being a familiar instance, taken up in detail in §6.3.[2] What this paper adds to Turing's diagnosis is formalisation: the Bayesian reconstruction immediately below, the Ockhamite analysis of §7.1, the diagnostic checklist of §2.2, and the case-by-case verdicts of Table 2. The structural insight is his.

4.1 The Turing Test and Its Discontents

A structural cost of Turing's mechanism-blindness emerges in the era of mechanistic interpretability. The test gives behavioural evidence a formal role while leaving mechanism evidence with no comparable channel into the verdict, and mechanism is therefore free to be deployed asymmetrically: ignored when behaviour is impressive, invoked when attribution is to be resisted.

[2]The miscount is an error of functioning at the tokenisation layer—the model does not have characters in its input representation—not an error of conclusion at the layer where reasoning is being assessed. Turing diagnosed the conflation seventy-five years before tokenisation existed.

The diagnosis suggests a natural *Bayesian reconstruction* of the imitation game. Posterior credence that a system is intelligent is the product of a likelihood (behavioural performance) and a prior built from mechanism evidence—architecture, training regime, internal representations recovered by interpretability methods, causal interventions on those representations, and characteristic failure modes—so that the verdict tracks the joint product rather than the likelihood alone. The point is not that mechanism defeats behaviour or that behaviour defeats mechanism, but that both should update attribution explicitly and symmetrically, mechanism evidence forced into the calculation rather than smuggled in case by case.

A clean *Gedankenexperiment* isolates the effect. Imagine the very same LLM arrived as a black box recovered from an extraterrestrial probe—same behaviour, same internal representations, same performance profile, only construction-history missing. Many readers would assign a higher prior in the counterfactual than they assign to the actual system. The *Gedankenexperiment* is not itself an argument that construction history is irrelevant: construction history is real evidence about a system's function class, and a skeptic who conditions the prior on it is not making an obviously inconsistent move. The substantive question is how much evidential work construction history can do once behavioural and mechanistic evidence accumulates.

On the simplest description, gradient descent produces systems that minimise predictive loss by exploiting statistical regularities. But the systems reviewed in §5 also exhibit recoverable world-state representations, high performance on some theory-of-mind operationalisations, and calibrated self-assessment under some conditions. The explanatory gap between training objective and cognitive performance is often called *emergence*. A structurally similar gap exists in biological cognition: we describe neurons and synapses, but we do not have a complete causal account of how they yield understanding, reasoning, or self-modeling. Granting emergence in the biological case while ruling it out for LLMs solely because we know their construction history requires an additional argument; it is not licensed merely by the fact that one system is evolved and the other engineered.

Read through the Bayesian reconstruction, the two canonical objections to the Turing test acquire a clean reformulation. Block's (1981) *Blockhead*—a giant lookup table that, by construction, produces every appropriate Turing-test response—is best read not as a counterexample to Turing but as the limiting case in which mechanism evidence overrides behavioural evidence: behavioural likelihood is maximal, mechanism inspection reveals pure lookup with no compression and no abstraction, and the prior conditioned on that mechanism collapses, taking the posterior with it. The argument's force is preserved, but its structure changes. Block was already, implicitly, doing what the reconstruction does explicitly—importing mechanism evidence to override behavioural evidence—and the reconstruction simply makes the move symmetric: mechanism can also raise the prior when it reveals world models, algorithmic circuits, or causally recoverable representations, as §5 documents. A further constraint discharges most of the residual worry: any *implementable* system passing a sufficiently rigorous Turing Test must be doing something more interesting than lookup (§5.1).

Searle's (1980) Chinese Room is the more enduring challenge, and the reconstruction allows it to be re-read without the Systems Reply doing all the work. The Chinese Room sets up high behavioural performance against a mechanism—symbol-manipulation following rules—that Searle stipulates cannot constitute understanding. The reconstruction asks the diagnostic question of the stipulation: what observable evidence could revise the verdict that this kind of mechanism does not constitute understanding? The argument does not specify; the prior is fixed at the level of construction history, not by anything the room does or could do. The Systems Reply, which Searle himself anticipated [Searle, 1980], captures one symptom of this—understanding may belong to the room as a whole even when no part of it understands anything—and the split-brain literature [Gazzaniga, 2005] has shown for half a century that the same systems-level structure characterises human cognition: no individual neuron or

brain compartment understands anything, yet understanding plausibly emerges from the integrated system. If component-level absence of understanding entailed system-level absence, human brains would not understand either, and the argument proves too much. Read through the reconstruction, the Chinese Room is an early instance of the mimicry-trap structure §2.2 diagnoses: a behaviour-conceding, mechanism-dismissing argument whose prior is fixed in advance of any observation.

The two objections, then, occupy different positions under the reconstruction: Block's is a legitimate Bayesian move with a responsive prior, Searle's a prior-level commitment that no observation could revise. The contemporary mimicry-skeptical arguments taken up next inherit Searle's structure rather than Block's, and the diagnostic of §2.2 catalogues the result.

4.2 Stochastic Parrots and Their Afterlife

The contemporary mimicry-skeptical arguments about LLMs share Searle's structural move applied to the system's empirical features. Each takes the descriptive fact that LLMs are, at some implementation layer, statistical pattern-completers, and slides from that fact to the normative claim that such a system cannot, in principle, understand or reason. The slide concedes behaviour but lets mechanism settle the verdict by stipulation: no observation about what the system does, or how its internal representations are organised, could in principle revise the conclusion. Two contemporary expressions of the move have particular currency in the LLM debate, and a third—Floridi's more sophisticated axiomatic version—is treated separately in §4.3.

The most influential statement is Bender et al.'s (2021) "On the Dangers of Stochastic Parrots" [Bender et al., 2021]. The original paper made several distinct claims, not all about intelligence attribution: environmental costs of training, encoded biases in models trained on internet text, and the risk that fluent text generation misleads users into attributing comprehension where none exists. These concerns are legitimate and the framing was a reasonable description of systems like BERT and GPT-2. This paper does not retrospectively criticise Bender et al. for failing to anticipate the trajectory of LLM capability. What it diagnoses is the persistence of the framing today: the continued description of LLMs as systems that "haphazardly stitch together" text "without any reference to meaning" (Bender et al. 2021, p. 617) increasingly diverges from the empirical record and has become structurally Searlean—behaviour conceded, mechanism settling the verdict, no specification of what evidence would revise it. A more measured contemporary statement of the cautionary line in linguistics is Futrell and Mahowald's [Futrell & Mahowald, 2025] BBS target article, which treats LLMs as model systems and proofs of concept for theoretical linguistics rather than as replacements for it—a position consistent with the empirical reading developed here while preserving the disciplinary stake the original "parrots" paper sought to defend.

The second contemporary expression comes in two variants of the same move, popular and technical. Chomsky, Roberts, and Watumull's (2023) *New York Times* essay argues that LLMs are "stuck in a prehuman or nonhuman phase of cognitive evolution" because they operate through statistical pattern-matching rather than rule-based grammar (more developed academic versions of the view—in generative grammar, symbol grounding, embodied cognition, and teleosemantics—are taken up in the steelman of §5.6). The technical variant is the *next-token reductio*: LLMs are *just* next-token predictors, and therefore cannot really reason, plan, or understand. The factual premise is unobjectionable in both forms; the conclusion is not. Biological brains, at the level of base operations, are equally "just" electrochemical signal-propagators, and no individual neuron reasons. If we are permitted to slide from "LLMs predict tokens" to "LLMs do not really reason," we are equally permitted to slide from "brains propagate action potentials" to "brains do not really reason"—which is absurd, and the inference is invalid in

both directions. A training objective is not the same thing as what the trained system does: a system optimised to predict tokens, given sufficient scale, may have to develop world models, algorithmic circuits, theory-of-mind capacities, and meta-cognitive markers in order to predict tokens well, and §5 documents this in catalogue form. The residual force of both moves comes from a level-of-description error: the description of a system at its lowest implementation layer is not a description of what it does at higher layers, and inferring the absence of cognition from the presence of mechanism is substrate prejudice in another form.

4.3 The Axiomatic Strategy: Floridi and "Agency Without Intelligence"

The arguments examined so far operate by dismissing demonstrated performance. A more sophisticated version of the trap operates not by denying performance but by defining intelligence in ways that exclude artificial systems by description. The work of Luciano Floridi is the clearest contemporary instance, and deserves a careful reconstruction rather than a paraphrase.

Floridi's central thesis, developed across [Floridi, 2023a, Floridi, 2023b] and extended in [Floridi, 2025a, Floridi et al., 2025], is that AI represents "agency without intelligence." The clearest single statement is in [Floridi, 2023a]: LLMs "can process texts with extraordinary success and often in a way that is indistinguishable from human output, while lacking any intelligence, understanding or cognitive ability." The 2025 follow-up adds that what consumers of LLM output experience is best understood as "semantic pareidolia," the perception of meaning where none is present [Floridi, 2025a]. The argument can be reconstructed from Floridi's own formulations:

1. *(P1) Definition.* Intelligence requires the processing of "mental content or meanings," i.e. genuine semantic engagement.
2. *(P2) Description.* LLMs operate "statistically, that is, working on the formal structure, and not on the meaning of the texts they process" [Floridi, 2023a].
3. *(C) Conclusion.* LLMs lack intelligence, however sophisticated their behavior may be.

The conclusion follows. The trouble is the joint between (P1) and (P2): once both are accepted, no empirical observation about LLM behavior can bear on (C), because (C) is a deductive consequence of (P1) and (P2) and not a hypothesis about any further fact. The argument *functions tautologically* unless (P2) is independently defended by criteria that could in principle be falsified by behavioral or mechanistic evidence. Without such independent defense, (P2) is not a substantive description that the data could revise; it is the conclusion (C) imported into the description of the candidate system. The 2025 categorical-analysis paper [Floridi et al., 2025], which argues that LLMs "circumvent" rather than solve the symbol-grounding problem, has the same structure: rigorous on its own terms, but its bearing on the question of LLM intelligence depends entirely on the prior commitment that intelligence requires the kind of grounding LLMs are described as circumventing.

Read through the Bayesian reconstruction of §4.1, the same diagnosis takes a sharper form. (P1) is not a likelihood claim but a *prior*, set definitionally to zero for any system Floridi calls an LLM. A posterior conditioned on a zero prior cannot be updated by any quantity of behavioural or mechanistic evidence, which is what Lindley's Cromwell rule [Lindley, 1985] forbids in any empirical context. The tautology diagnosed above is in this sense Floridi's framing violating Cromwell's rule by construction: the prior on the proposition the data could update has been fixed at a value no data can reach. The full Bayesian formalisation of this failure mode—and its relation to the parsimony argument that closes the paper—is developed in §7.1.

Three features make the diagnosis precise. First, the *unfalsifiability*: what conceivable behavioral evidence could show, on Floridi's framing, that an LLM processes meanings rather than merely

formal structures? Sophisticated correct outputs are filed under sophisticated statistical processing; articulated reasoning is "simulated" reasoning; behavioral tests are "semantic pareidolia." No specification is offered, anywhere in the corpus, of an observation whose presence would falsify (P2) or move the verdict on (C). Second, the *temporal displacement* (cf. Gahrn-Andersen, 2025): the original 2023 paper cited LLM "brittleness," failures at "simple mathematics," and inability to pass the Turing Test as the empirical bases of the verdict; several have since been substantially addressed (§5), and the criteria have moved from performance to mechanism rather than the verdict updating. Third, the *burden of (P2) is heavier than it looks*: it is not the uncontested claim that LLMs are implemented through statistical learning, but the much stronger claim that the resulting computation is exhausted by formal-structure manipulation with no engagement of meaning. As §5 documents, the function class implemented by these systems demonstrably includes recoverable internal world models, algorithmic circuits, and abstract conceptual features; whether any of this constitutes engagement with meaning depends on what one means by "meaning," which is precisely the question (P1) was supposed to settle.

This is not to dismiss Floridi's concerns. Questions about grounding, about the relationship between statistical patterns and meaning, about the differences between human and artificial cognition are legitimate, and the steelman of §5.6 treats several of them as live. The diagnosis offered here is also procedurally fair: if Floridi (or a Floridi-defender) supplies independent, falsifiable criteria for what would count as semantic engagement on the part of an LLM—criteria operational enough that some conceivable behavioral or mechanistic finding could establish or deny their satisfaction—then the charge of tautology should be withdrawn, and the dispute becomes a substantive empirical or philosophical disagreement about those criteria rather than a mimicry-trap diagnosis. The narrower point being made here is methodological: as the argument is currently built, the axiomatic strategy forecloses inquiry into the questions it appears to be answering. The mimicry trap, in its most sophisticated form, does not deny that the parrot speaks. It redefines "speaking" to exclude whatever the parrot does.

5 Empirical Challenges to the "Mere Statistics" View

Several of the critiques diagnosed in §4—especially the crude "parrot" framing and the next-token reductio—predict or imply that LLM behavior is exhausted by surface co-occurrence statistics. The empirical record now constrains what these systems do internally and makes that view increasingly difficult to sustain. This section supplies the empirical content that three of the diagnostic tests from Table 1 require in order to fire. It is the candidate evidence the *Falsifiability* test asks the mimicry-skeptic to engage with; it speaks to the *Mechanism* test by showing what mechanism inspection actually reveals (world models, algorithmic circuits, calibrated self-assessment, causally manipulable abstract features); and it constrains the *Contamination* test by characterising the function class implemented by these systems rather than treating training-data overlap as a wholesale defeater. Five lines of evidence are reviewed (compression, world-model emergence, mechanistic interpretability, mathematical reasoning, and theory-of-mind / meta-cognition), and the strongest remaining skeptical position is then put on the table (§5.6). §6 takes up the *Consistency* test that has so far prevented this evidence from registering.

5.1 The Compression Argument

A capable open-weight model such as Llama 3.1 70B contains roughly 70 billion parameters, requiring approximately 140 GB of storage. Its training corpus comprises trillions of tokens, conservatively 4–15 TB of raw text. The model is thus 30 to 100 times *smaller* than its training data, and so cannot store that data; it must compress it. The point is well established in the public ML discourse and not original to this paper. What is worth pinning down, because

the argument is often pushed further than it can support, is its scope. *Global* lookup-table memorisation is what the parameter-count argument rules out. *Local* memorisation remains a live concern: models can and demonstrably do memorise specific benchmark items, rare facts, and high-value passages, and contamination at this level is a methodological worry empirical work must address rather than dismiss. Compression is also necessary but not sufficient for understanding: gzip exploits redundancy without comprehending anything. What the argument does establish is the narrower constraint that any account of these systems must explain how something too small to store its training data nevertheless reproduces structure from it across novel inputs. The kind of abstraction this requires is the empirical question the rest of the section takes up.

5.2 Emergent World Models

Li et al. [Li et al., 2022] trained a language model on Othello move sequences alone, with no representation of board state. Probing the trained network recovered an accurate internal model of the board, never explicitly provided in training. Karvonen [Karvonen, 2024] extended the approach to chess: a 50M-parameter model trained on PGN move strings supports linear probes that recover the current board state, including compositional features such as check, castling rights, and pinned pieces, plus a manipulable latent representation of player skill. Gurnee and Tegmark [Gurnee & Tegmark, 2024] probed Llama-2 family models for representations of geographic and temporal entities and recovered metric coordinates: latitude and longitude of cities, dates of historical events, relative positions of buildings within a city, none of which the model ever saw as maps or timelines. At the largest scale yet investigated, Templeton et al. [Templeton et al., 2024] used sparse autoencoders to decompose Claude 3 Sonnet into approximately 34 million interpretable features ranging from entity-level concepts (the Golden Gate Bridge) to abstract structures (deception, sycophancy); the features are causally manipulable, with clamping producing predictable behavioral shifts.

A skeptical reading is available: a system optimised for next-token prediction may develop such latents because they are predictively useful, and predictive utility is not yet semantic understanding. That reading is fair, but it falsifies the simplest surface-statistics accounts on which LLM behavior is exhausted by surface co-occurrence. The probing results raise the evidential cost of the strong mimicry hypothesis without discharging the harder semantic-grounding debate, which is taken up in the steelman of §5.6 and the parsimony argument of §7.1.

5.3 Mechanistic Interpretability

Mechanistic interpretability has documented increasingly sophisticated computational structures inside neural networks. Nanda et al. [Nanda et al., 2023] identified circuits in transformer models that implement modular arithmetic through Fourier-basis representations and trigonometric identities (a small algorithm rather than a lookup table), and tracked the transition from memorisation to algorithm during training—the so-called *grokking* phenomenon. Elhage et al. [Elhage et al., 2021] documented "induction heads" that implement a general pattern-completion algorithm explaining a substantial fraction of in-context learning. Wang et al. [Wang et al., 2022] reverse-engineered the GPT-2 small circuit for indirect object identification: twenty-six attention heads in seven functional classes (name-mover heads, S-inhibition heads, duplicate-token heads) compose into a multi-step computational graph. Arditi et al. [Arditi et al., 2024] showed that refusal behavior across thirteen safety-tuned LLMs is mediated by a single one-dimensional residual-stream subspace, causally manipulable in either direction. Fraser-Taliente et al. [Fraser-Taliente et al., 2026] trained natural-language autoencoders

to produce unsupervised text descriptions of arbitrary activation vectors; the descriptions are causally valid (steering vectors derived from them alter behavior as predicted) and surface representational content the model does not itself verbalise—including a form of "unverbalised evaluation awareness," where the system internally represents the suspicion of being evaluated without stating it.

Mechanistic interpretability is a young field whose specific methods (linear probes, sparse autoencoder features, circuit attribution) are themselves contested, and the findings above should be read as evidence rather than settled fact. With that caveat, the cumulative picture is robust enough to bear the argumentative weight placed on it here: a flat landscape of memorisation circuits would have been a vindicating finding for the strong mimicry account, and the field is finding structured algorithmic implementation, manipulable feature directions corresponding to abstract concepts, and behavioral circuits with the architecture of decision systems instead. The claim is not that this settles semantics—the steelman of §5.6 treats that as open—but that it raises the evidential cost of the pure mimicry hypothesis.

5.4 Mathematical Reasoning and the Contamination Question

The data-contamination objection holds that an LLM's apparent capability may be retrieval rather than reasoning, because the relevant solutions appeared in training data. The objection has two very different uses. *Legitimate contamination control*—held-out test sets, post-training benchmarks, formal verification, canary strings, dataset audits, contamination probes, synthetic problem generation—is a routine part of LLM evaluation, and an evaluation that has not addressed it cannot speak to capability. *Blanket contamination dismissal* is contamination as an in-principle defeater, deployed wholesale regardless of whether the specific evaluation has controlled for overlap, and without advance specification of what would discharge it. The blanket version proves too much: human cognition is built on "training data" (experience, education, cultural exposure), and we do not demand of human mathematicians proof that they never encountered a similar problem during their education—encountering related problems is how one *becomes* capable of mathematical reasoning. The legitimate version asks the question that should be asked: how much of this performance is overlap, and how would we know?

Mathematics is a methodologically clean test bed for that question, because formal verification narrows the contamination defense progressively. Feng et al. (2026) deployed a mathematics research agent built on Gemini Deep Think over 700 open conjectures from Paul Erdős's problem database. The system produced apparently novel solutions to five problems that human mathematicians verified as correct, with one subsequently formalized in the Lean proof assistant. The authors handle the contamination concern explicitly and repeatedly, scanning reasoning traces to bound the risk of "subconscious plagiarism"—a model instance of the legitimate contamination control above. Google DeepMind's Gemini Deep Think achieved gold-medal performance on the 2024 IMO [Google DeepMind, 2025], solving five of six problems with proofs verified by mathematicians. The contamination defense applies most easily to historical benchmarks, less easily to formally checked solutions, and least easily to genuinely novel post-training problems whose solutions are mechanically verified.

5.5 Theory of Mind and Meta-Cognition

Theory of mind, the capacity to attribute mental states to others, is a textbook "real intelligence" criterion whose place in the contemporary AI debate parallels its place in animal cognition (§3.1). Strachan et al. [Strachan et al., 2024] tested GPT-4 and Llama-2 on a battery of theory-of-mind tasks against 1,907 human participants. GPT-4 matched or exceeded human performance on some operationalisations (indirect requests, false beliefs, misdirection); underperformance on

faux-pas detection traced to over-conservative answering rather than failure of inference. The strict claim is that behavioral performance on these task families is at or above the human-non-expert level for some operationalisations, with task-specific deviations whose mechanisms can be characterized; whether that constitutes possession of theory of mind is a further question (§2). Meta-cognitive markers are also accumulating: Kadavath et al. [Kadavath et al., 2022] showed that LLMs predict the probability that their own answers are correct with reasonable calibration, improving with scale and contextual prompting. Each finding is open to challenge as "mere imitation," and well-known confounds (prompt sensitivity, format-specific heuristics, evaluation leakage) must be addressed for any specific result. As scale, robustness, and out-of-distribution transfer increase, the burden rises on the claim that such behavior is merely an imitation of meta-cognition rather than a functional analogue of it; an argument we return to in §7.

5.6 The Strongest Case for Mimicry-Skepticism

The empirical record above puts pressure on crude mimicry accounts of LLM behavior, but it does not refute every form of skepticism, and it would be self-serving to leave that impression. The arguments diagnosed in §4 are the versions of mimicry-skepticism that most clearly exhibit the trap; several stronger skeptical positions accept the empirical record just reviewed and locate their disagreement elsewhere. These are not objections to be dismissed: they are the strongest candidates for turning mimicry-skepticism into an evidence-responsive research program, and the diagnostic apparatus of this paper is meant to separate them from their evidence-resistant cousins, not to suppress them. The strong skeptic, on the reading offered here, accepts the findings of §5.1–§5.5 and the level-of-description point, concedes that crude mimicry accounts are no longer viable, and nevertheless denies that LLMs possess intelligence in the sense that matters, on one or more of the following grounds.

1. *Grounding.* The symbols an LLM manipulates are not causally connected to perception and action; whatever internal structure they develop does not anchor in the sensorimotor history that, on causal theories of reference [Gahrn-Andersen, 2025], makes a representation a representation *of* something.

2. *Agency.* A base LLM has no endogenous goals, no persistent project, no continuing identity across sessions. Apparent goal-directedness is scaffolded by prompt and deployment.

3. *Normativity.* Language use is governed by norms of correctness the speaker can recognize as norms and revise; an LLM produces norm-conforming output statistically without standing in the right relation to the norms themselves.

4. *Embodiment.* Human meaning is shaped by bodily interaction with a physical environment; a system that has never had a body cannot mean what we mean.

5. *Social-pragmatic.* Linguistic understanding is constituted by participation in social practices (requests, promises, corrections, accountability), not observation of their textual residue.

6. *Training-objective character.* Next-token prediction over human discourse may produce a simulator of discourse rather than an agent who discourses.

7. *Evaluation.* Benchmarks reward fluency, plausibility, and test-taking competence; high scores may track these properties rather than the intelligence the benchmarks were intended to measure.

These objections have very different epistemic statuses, and the diagnostic apparatus of this paper applies to some and not others. Objections (4) embodiment and (5) social-pragmatic are *legitimate but largely out of scope*: they concern conditions on meaning, agency, or selfhood the paper does not contest and is not in a position to settle. Held honestly as theses about those conditions rather than as denials of functional capacity, they are not in the trap. Objections (1) grounding, (3) normativity, and (6) training-objective character are *relevant and inconclusive*: they specify properties at least in principle empirically tractable (grounding via multimodal training and world-model recoverability; normativity via self-correction and response to social signals; training-objective effects via what behaviours actually emerge), and the current empirical record speaks to all three without settling them. A skeptic who treats these as live empirical questions and specifies what would update them is doing science. Objections (2) agency and (7) evaluation are *partially absorbed*: scaffolded agency is, as §2 acknowledged, weaker than endogenous agency, and benchmark-tracking is a real phenomenon; held as in-principle defeaters that no system or measurement could discharge, the trap diagnosis applies, but as ongoing critiques of specific architectures and measurement methods they are part of the empirical conversation this paper takes itself to be inside.

The mimicry trap is therefore not the claim that all skepticism about LLM intelligence is bankrupt. It is the narrower claim that a recurring form of AI skepticism is methodologically defective because it cannot say what evidence would change its mind. The strongest reply to the diagnosis is not to deny the empirical record wholesale, but to specify which additional property is required, why it belongs to *intelligence* rather than to consciousness, agency, embodiment, or moral status, and what empirical signature would indicate its presence or absence. A skeptic who can do that is not in the trap; the diagnostic apparatus is designed precisely to pick out the cases where they cannot.

6 The Epistemological Double Standard

The mimicry trap's deepest manifestation is not in any specific argument but in the *asymmetric application of standards*. When humans exhibit intelligent behavior, we infer intelligence; when AI systems exhibit identical behavior, we demand additional proof, and the required proof is systematically unspecifiable. This section is the diagnostic's *Consistency* test (Table 1) applied at length. Consistency is a symptomatic test, and symptomatic tests are diagnostic only when the asymmetry is demonstrated as a pattern across cases rather than as a single instance of legitimate skepticism: any one case can be discussed away on its merits, but the accumulation of cases is the diagnosis. What follows is therefore a survey, not a polemic—each individual case admits substantive debate, while their joint pattern is what the diagnostic registers.

6.1 Three Asymmetries

The asymmetry takes three reinforcing forms. First, *transparency bias*: because we increasingly understand how neural networks function, their intelligence seems reducible to "mere" calculation, while the brain's comparatively opaque mechanisms get mystified into evidence of special status. Transparency should increase our confidence that a system is doing something interesting, not decrease it. Second, *substrate prejudice*: when a biological system exhibits flexible, goal-directed behavior we attribute intelligence without demanding proof of any particular internal process (octopus intelligence is distributed across eight arms; corvid intelligence operates through an avian pallium with no homology to mammalian cortex), but when the substrate is silicon, additional qualities (real understanding, genuine reasoning) are demanded that behavioral evidence cannot establish. The same asymmetry has been articulated at the level of vocabulary by Shanahan [Shanahan, 2024a, Shanahan, 2024b]: mentalistic terms ("the model

believes," "the model knows") are said to mislead about silicon systems even though we apply them routinely to biological systems whose mechanism is no better understood. Third, *retreat to unfalsifiability*: when benchmarks are met they are dismissed as "mere" task completion; when capabilities are demonstrated they are attributed to contamination; when internal representations are documented they are dismissed as not really understanding. At each stage the critic specifies new requirements without articulating what evidence would suffice.

6.2 The Shifting Threshold: Turing Test as Canonical Instance

The AI Effect is documented across many benchmarks; the Turing Test is the canonical case. "The Turing Test" is not a single fixed protocol but a family of imitation-game operationalisations, several of which have now been passed under controlled conditions: Jones and Bergen [Jones & Bergen, 2024] found GPT-4 judged human 54% of the time, indistinguishable from chance; a subsequent study [Jones et al., 2025] found GPT-4.5, when appropriately prompted, achieved human-indistinguishable performance in the original three-party imitation game at rates significantly above chance. Many philosophers, of course, had rejected the Turing Test as sufficient long before these results, so the response is not simply "acceptance until LLMs passed it." What is instructive is the response from those who had treated Turing-style indistinguishability as a meaningful operational threshold: rather than treat the results as strong evidence, the test was widely redescribed as measuring only deception, fluency, or imitation. A similar pattern appears in chess after Deep Blue, in Go after AlphaGo, in professional licensing examinations once LLMs began passing them, and in mathematical olympiads after Gemini's gold-medal IMO performance [Google DeepMind, 2025]; each had functioned, in some relevant discourse, as a salient marker of intelligence or advanced cognition until systems met it, after which it was dismissed as "mere" something.

6.3 The Error Asymmetry: Strawberries and Architecture

A briefer manifestation concerns the treatment of errors. LLM mistakes, even trivial ones, are routinely treated as decisive evidence against intelligence; equivalent human mistakes are dismissed as incidental. The "strawberry" example is emblematic: early LLMs, asked how many times "r" appears in "strawberry," frequently answered "2." This is not decisive evidence of absent understanding; it is at least partly explicable as architectural, because LLMs operate on multi-character subword tokens and character-level frequencies are not directly available. Humans have analogous architectural limitations (we cannot instantly compute 347×892, reliably count syllables while speaking, or override the conjunction fallacy even when we know the correct answer). The asymmetry is not in the existence of architectural limits but in whether they are read as evidence about cognition or as ordinary signatures of a constrained processing substrate.

6.4 When Superior Performance Becomes "Illusion"

A compact contemporary instance can be drawn from Loru et al.'s (2025) PNAS paper "The Simulation of Judgment in LLMs" [Loru et al., 2025]. Six LLMs were compared to human non-experts on a news-source credibility task benchmarked against NewsGuard expert ratings. LLMs achieved 85–97% agreement with the experts; humans performed at chance. The paper nonetheless framed the LLMs' *superior* performance as "epistemia," the "illusion of knowledge emerging when surface plausibility replaces verification," on the grounds that the models "rely on lexical associations and statistical priors rather than contextual reasoning." Human failure was characterized as the use of "different and less consistent indicators"—a methodological dif-

ference, not a deficit. The asymmetry is the diagnostic feature: identical-or-better performance by the silicon system is pathologised as simulation; chance-level performance by the carbon system is naturalised as alternative method. The most sophisticated form of the asymmetric pattern, which moves from dismissing performance to defining intelligence so that performance cannot in principle constitute evidence of it, has already been treated in §4.3 as the culminating case of the contemporary mimicry-formulations.

7 Implications and Conclusions

7.1 The Parsimony Argument: *Entia non sunt multiplicanda*

The Turing test posed the question this paper has been pursuing: if a system's behavior is indistinguishable from that of an intelligent agent across all observable dimensions, what justifies positing an invisible absence? The venerable answer is *Entia non sunt multiplicanda praeter necessitatem*: entities are not to be multiplied beyond necessity. The maxim, conventionally attributed to William of Ockham and crystallised by his fourteenth-century successors [Spade, 1999], targets a specific intellectual sin: positing *entia*, invisible items in one's ontology, whose only function is to preserve a conclusion the evidence does not require.

This is the standard the mimicry-trap argument fails. The grammar of "mimicry" makes it visible: to say that a system mimics intelligence is to assert a relation, that it produces outputs resembling those of a real intelligent agent but lacks the underlying property P that would make those outputs the genuine article. Without P, "mimicry" collapses into "does the same thing as," which is not mimicry but identity. The mimicry framing therefore requires its user to specify, at least implicitly, what P is. The same structure recurs across the qualifier vocabulary of this debate—*real* intelligence, *true* semantic content, *genuine* understanding, *authentic* reasoning. In each case the qualifier marks the position of an unobservable property whose only function is to license the negative verdict; phlogiston, ether, and vital force are the canonical instances of the same structure. The irony is that mimicry sceptics frequently invoke parsimony against attributions of cognition ("don't anthropomorphise"); their own account is the one multiplying *entia*.

The structural form can be made explicit. Let I denote the proposition that a system is functionally intelligent in the sense of §2; B the behavioral evidence; M the mechanism evidence. Bayes' rule gives $P(I \mid B, M) \propto P(B, M \mid I) \cdot P(I)$. The mimicry-skeptical account introduces a further variable E, an unobservable "essence" such that intelligence proper obtains exactly when E does, but with no observable consequences for B or M. The likelihood $P(B, M \mid I, E)$ is, by stipulation, independent of E: the same data are predicted whether E holds or not, so E is *non-identifiable* and the inference about whether the system is "really" intelligent is the inference one was always going to make. Non-identifiability is the formal expression of Ockham's intuition: the question is not whether a posit is observable, but whether it makes a difference to expectation.

A complementary failure operates at the level of the prior rather than the likelihood. If a skeptic enters with $P(I) = 0$ for any system whose substrate disqualifies it antecedently, then by Bayes' rule $P(I \mid B, M) = 0$ regardless of B or M: no behavioral or mechanistic evidence, however strong, can shift the verdict. Lindley named this pathology *Cromwell's rule* [Lindley, 1985], after Oliver Cromwell's 1650 letter to the General Assembly of the Church of Scotland: "I beseech you, in the bowels of Christ, think it possible that you may be mistaken." Non-identifiability and the Cromwell violation diagnose distinct pathologies—the first locates the failure in an evidentially empty posit, the second in a prior immune to revision—but they often co-occur, and most mimicry-skeptical arguments in practice combine an unfalsifiable essence with a

substrate-bound prior. The falsifiability test of §2.2 is the operational form of asking whether the skeptic's prior in fact lies strictly between 0 and 1.

This must be distinguished from the legitimate unobservables science routinely employs. Electrons, genes, latent psychometric variables, dark matter, and mental states are all unobservable; Ockham's razor does not cut them because each has independent evidential consequences (charged-particle behavior, inheritance distributions, response-time profiles, gravitational lensing) that connect it to more than the single fact it was introduced to explain. The problem with the mimicry-skeptic's E is not invisibility but the absence of independent evidential consequences. A skeptic could escape the diagnosis by specifying, in advance, what behavioral or mechanistic evidence would distinguish a system that has E from one that lacks it; that skeptic is doing science. The diagnosis applies to the version in which E is invoked precisely because it cannot be operationalised, and is preserved precisely because it cannot.

This is also the precise diagnosis of the tautology pattern this paper has tracked. If intelligence is defined functionally, then a system exhibiting all the functional markers is, by that definition, intelligent. To insist that it nevertheless lacks "real" intelligence is to redefine the term mid-argument. Jefferson's "produced a Phillis Wheatley; but it could not produce a poet" is the eighteenth-century instance; "stochastic parrot" is its contemporary descendant. The Ockhamite reading and the tautology diagnosis coincide. The empirical evidence reviewed in §5 is what makes this argument bite: given what is now known about LLM internal representations, the gap that low-complexity mimicry would produce has not been observed, and the residual skeptical work is being done by an entity the data cannot see. The argument concerns intelligence, not phenomenology: the question of consciousness, sentience, or moral status remains open.

7.2 The Default Has Shifted

The Ockhamite argument has a methodological corollary. For most of the history of artificial intelligence the default null hypothesis was reasonable: artificial systems lacked the markers of intelligence; the parsimonious starting point was that intelligence was absent, and demonstrations had to overcome this presumption. That context no longer obtains. The cumulative evidence reviewed in §5—world-model emergence, algorithmic circuits, sparse-feature decomposition, high performance on theory-of-mind operationalisations, calibrated self-assessment under some conditions, novel mathematical proofs verified mechanically—does not prove that LLMs are intelligent. What it does is dissolve the empirical situation that made the original prior reasonable. The default has shifted not only because behavior has improved but because mechanism has become inspectable, and inspection has not revealed the lookup tables the original prior anticipated.

The diagnosis applies forcefully to one form of agnosticism and not to another. *Undisciplined agnosticism*, which declines to update regardless of evidence and offers no advance specification of what would suffice for either attribution or denial, preserves the inherited prior under the cover of caution, and the burden-shift argument applies to it directly. *Disciplined agnosticism*, which acknowledges the cumulative evidence, distinguishes target concepts (functional intelligence vs. richer cognition, semantic grounding, autonomous agency), and remains undecided about a specifically named question while specifying what would update its position, is not in the trap. A reader who, having considered the evidence, accepts that current systems exhibit functional intelligence in the sense of §2 but withholds judgment on grounded semantics or fuller agency is not failing to update; they are localising their uncertainty.

The defensible version of the burden-shift claim is therefore narrower than its first formulations. Blanket denial of LLM intelligence simpliciter no longer enjoys the inherited prior; it is now a substantive empirical claim that owes its own evidence, and the cumulative pattern

of evidence-resistant moves documented earlier (benchmarks dismissed once met, internal-representation findings reinterpreted, functional markers downgraded) is increasingly insulated from disconfirmation. The methodological consequence is that Turing-test-style thresholding is no longer the only or central apparatus we need; the apparatus appropriate to the new question is dispositional rather than confirmatory: not "has the system crossed the threshold?" but "what failures would warrant revising the default attribution downward?" This is closer to how comparative cognition characterises non-human cognition.

7.3 The Diagnostic Applied

The diagnostic of §2.2 has been deployed implicitly throughout the paper. Table 2 consolidates the verdicts on the cases reviewed: the three historical instances of §3, the four contemporary mimicry-formulations of §4, the Loru et al. "epistemia" case from §6.4, and the blanket version of the contamination objection treated in §5.4.

The historical cases fire the structural tier completely, and their registration is now uncontroversial: the scientific and moral communities that held the positions have themselves updated. Of the contemporary LLM-skeptical formulations, the persistent stochastic-parrots framing, the next-token reductio, and Chomsky's mechanism critique fire the same structural pattern as the historical cases, with mechanism dismissal as the dominant symptomatic move and asymmetric standards across substrates as a consistent secondary pattern. Floridi's axiomatic strategy is the maximal case: its definitional structure makes the unfalsifiability, the asymmetric standards, and the goal-post displacement most explicitly visible, because the move from dismissing performance to defining intelligence so that performance cannot in principle constitute evidence of it leaves nothing implicit. The blanket version of the contamination objection fires a complementary symptomatic pattern: it does not appeal to an invisible essence or derive its conclusion from the system's nature, but constructs an in-principle defeater out of the system's training history—with the corresponding asymmetry to human cognition (which is also built on "training data") as its diagnostic feature.

The diagnostic registers the contemporary LLM-skeptical positions in the same family as the historical errors. That registration is the substantive output of the framework: not a rebuttal of every form of LLM skepticism (the steelman of §5.6 lays out which positions remain in good standing), but an empirical claim about which versions are operating in the trap.

		Structural			*Symptomatic*		
Case	Falsifiability	Invisible Absence	Ontological Precedence	Consistency	Goal-Post	Mechanism	Contamination
Jefferson on Wheatley	✓	✓	✓	✓	✓	—	✓
Corvid skepticism	✓	✓	(✓)	(✓)	✓	✓	—
Bee skepticism	✓	✓	✓	(✓)	✓	✓	—
Stochastic parrots (persistent)	✓	✓	(✓)	(✓)	✓	✓	—
Chomsky's mechanism critique	✓	✓	✓	✓	—	✓	—
Next-token reductio	✓	✓	✓	✓	—	✓	—
Floridi's axiomatic strategy	✓	✓	✓	✓	✓	✓	—
Loru et al. ("epistemia")	✓	✓	(✓)	✓	✓	✓	—
Blanket contamination dismissal	✓	—	—	✓	✓	✓	✓

Table 2: Application of the diagnostic to the cases reviewed in the paper. ✓ = clean fire; (✓) = partial; — = does not apply.

7.4 Summary of Contributions

The paper claims four contributions. *First*, the cross-domain consolidation: comparative cognition's diagnosis of anthropodenial, the AI Effect, and the Wheatley structural template are, on the analysis presented here, the same diagnosis applied to different substrates. *Second*, the diagnosis of certain influential definitional arguments as functioning tautologically: Floridi's "agency without intelligence" thesis (§4.3) is shown to deliver its conclusion of non-intelligence by the definitions used unless premise (P2) is independently defended by criteria that could in principle be falsified. *Third*, the default-shift argument: blanket denial of LLM intelligence simpliciter no longer enjoys the inherited prior, while disciplined agnosticism about specific richer notions remains in good standing. *Fourth*, the Ockhamite formulation: the mimicry-skeptical account multiplies invisible *entia* that have no operational definition and no explanatory function beyond preserving a predetermined verdict. The diagnostic checklist of §2.2, applied case by case in Table 2, consolidates these moves into a single applicable instrument.

Acknowledgements

This paper develops from a talk given at the ChemAI conference in Amsterdam in November 2025. I thank Marco Tibaldi for the invitation, and Giulio Dalla Riva and Stefano Bagnara for valuable comments on earlier drafts.

Funding statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests

The author declares none.

References

[Arditi et al., 2024] Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., & Nanda, N. (2024). Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems 38 (NeurIPS 2024)*. DOI: 10.48550/arXiv.2406.11717

[Bender et al., 2021] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. DOI: 10.1145/3442188.3445922

[Block, 1981] Block, N. (1981). Psychologism and behaviorism. *The Philosophical Review*, 90(1), 5–43. DOI: 10.2307/2184371

[Buckner, 2013] Buckner, C. (2013). Morgan's Canon, meet Hume's Dictum: Avoiding anthropofabulation in cross-species comparisons. *Biology & Philosophy*, 28(5), 853–871. DOI: 10.1007/s10539-013-9376-0

[Chittka, 2022] Chittka, L. (2022). *The mind of a bee*. Princeton University Press.

[Chollet, 2019] Chollet, F. (2019). On the measure of intelligence. *arXiv preprint arXiv:1911.01547*. DOI: 10.48550/arXiv.1911.01547

[Chomsky et al., 2023] Chomsky, N., Roberts, I., & Watumull, J. (2023, March 8). The false promise of ChatGPT. *The New York Times*.

[Dennett, 1991] Dennett, D. C. (1991). *Consciousness explained*. Little, Brown and Company.

[de Waal, 1999] de Waal, F. B. M. (1999). Anthropomorphism and anthropodenial: Consistency in our thinking about humans and other animals. *Philosophical Topics*, 27(1), 255–280. DOI: 10.5840/philtopics199927122

[de Waal, 2016] de Waal, F. B. M. (2016). *Are we smart enough to know how smart animals are?* W. W. Norton & Company.

[Elhage et al., 2021] Elhage, N., Nanda, N., Olsson, C., et al. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*. https://transformer-circuits.pub/2021/framework/

[Feng et al., 2026] Feng, T., Trinh, T., Bingham, G., et al. (2026). Semi-autonomous mathematics discovery with Gemini: A case study on the Erdős problems. *arXiv preprint arXiv:2601.22401*.

[Fraser-Taliente et al., 2026] Fraser-Taliente, K., Kantamneni, S., Ong, E., et al. (2026). Natural language autoencoders produce unsupervised explanations of LLM activations. *Transformer Circuits Thread*. https://transformer-circuits.pub/2026/nla/

[Floridi, 2023a] Floridi, L. (2023). AI as agency without intelligence: On ChatGPT, large language models, and other generative models. *Philosophy & Technology*, 36, 15. DOI: 10.1007/s13347-023-00621-y

[Floridi, 2023b] Floridi, L. (2023). *The ethics of artificial intelligence: Principles, challenges, and opportunities*. Oxford University Press. DOI: 10.1093/oso/9780198883098.001.0001

[Floridi, 2025a] Floridi, L. (2025). AI and semantic pareidolia: When we see consciousness where there is none. *Harvard Business Review Italia*.

[Floridi et al., 2025] Floridi, L., Jia, Y., & Tohmé, F. (2025). A categorical analysis of large language models and why LLMs circumvent the symbol grounding problem. *arXiv preprint arXiv:2512.09117*. DOI: 10.48550/arXiv.2512.09117

[Futrell & Mahowald, 2025] Futrell, R., & Mahowald, K. (2025). How linguistics learned to stop worrying and love the language models. *Behavioral and Brain Sciences*. DOI: 10.1017/S0140525X2510112X

[Gahrn-Andersen, 2025] Gahrn-Andersen, R. (2025). Beyond symbol processing: The embodied limits of LLMs and the gap between AI and human cognition. *AI & Society*, 40, 3105–3107. DOI: 10.1007/s00146-025-02382-y

[Gates, 2003] Gates, H. L., Jr. (2003). *The trials of Phillis Wheatley: America's first Black poet and her encounters with the Founding Fathers*. Basic Civitas Books.

[Gazzaniga, 2005] Gazzaniga, M. S. (2005). Forty-five years of split-brain research and still going strong. *Nature Reviews Neuroscience*, 6(8), 653–659. DOI: 10.1038/nrn1723

[Godfrey-Smith, 2016] Godfrey-Smith, P. (2016). *Other minds: The octopus, the sea, and the deep origins of consciousness*. Farrar, Straus and Giroux.

[Google DeepMind, 2025] Google DeepMind. (2025). Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad. *Google DeepMind Blog*.

[Gurnee & Tegmark, 2024] Gurnee, W., & Tegmark, M. (2024). Language models represent space and time. *Proceedings of the International Conference on Learning Representations (ICLR 2024)*. DOI: 10.48550/arXiv.2310.02207

[Hofstadter, 2007] Hofstadter, D. R. (2007). *I am a strange loop*. Basic Books.

[Jefferson, 1785] Jefferson, T. (1785). *Notes on the state of Virginia*. [Privately printed, Paris].

[Jones & Bergen, 2024] Jones, C. R., & Bergen, B. K. (2024). Does GPT-4 pass the Turing test? *Proceedings of NAACL-HLT 2024*, 5183–5210. DOI: 10.18653/v1/2024.naacl-long.290

[Jones et al., 2025] Jones, C. R., Rathi, I., Taylor, S., & Bergen, B. K. (2025). People cannot distinguish GPT-4 from a human in a Turing test. *Proceedings of FAccT 2025*. DOI: 10.1145/3715275.3732108

[Kadavath et al., 2022] Kadavath, S., Conerly, T., Askell, A., et al. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*. DOI: 10.48550/arXiv.2207.05221

[Karvonen, 2024] Karvonen, A. (2024). Emergent world models and latent variable estimation in chess-playing language models. *Conference on Language Modeling (COLM 2024)*. DOI: 10.48550/arXiv.2403.15498

[Katz et al., 2024] Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270), 20230254. DOI: 10.1098/rsta.2023.0254

[Legg & Hutter, 2007] Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17(4), 391–444. DOI: 10.1007/s11023-007-9079-x

[Li et al., 2022] Li, K., Hopkins, A. K., Bau, D., et al. (2022). Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*. DOI: 10.48550/arXiv.2210.13382

[Lindley, 1985] Lindley, D. V. (1985). *Making Decisions* (2nd ed.). London: John Wiley & Sons.

[Loru et al., 2025] Loru, E., Nudo, J., Di Marco, N., et al. (2025). The simulation of judgment in LLMs. *Proceedings of the National Academy of Sciences*, 122(42), e2518443122. DOI: 10.1073/pnas.2518443122

[Loukola et al., 2017] Loukola, O. J., Perry, C. J., Coscos, L., & Chittka, L. (2017). Bumblebees show cognitive flexibility by improving on an observed complex behavior. *Science*, 355(6327), 833–836. DOI: 10.1126/science.aag2360

[McCorduck, 2004] McCorduck, P. (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence* (2nd ed.). A. K. Peters.

[Mitchell & Krakauer, 2023] Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13), e2215907120. DOI: 10.1073/pnas.2215907120

[Nanda et al., 2023] Nanda, N., Chan, L., Liberum, T., Smith, J., & Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*. DOI: 10.48550/arXiv.2301.05217

[Schwitzgebel & Garza, 2015] Schwitzgebel, E., & Garza, M. (2015). A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*, 39(1), 98–119. DOI: 10.1111/misp.12032

[Searle, 1980] Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424. DOI: 10.1017/S0140525X00005756

[Seth, 2025] Seth, A. K. (2025). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*. DOI: 10.1017/S0140525X25000032

[Shanahan, 2024a] Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, 67(2), 68–79. DOI: 10.1145/3624724

[Shanahan, 2024b] Shanahan, M. (2024). Simulacra as conscious exotica. *Inquiry*. DOI: 10.1080/0020174X.2024.2434860

[Shevlin & Halina, 2019] Shevlin, H., & Halina, M. (2019). Apply rich psychological terms in AI with care. *Nature Machine Intelligence*, 1(4), 165–167. DOI: 10.1038/s42256-019-0039-y

[Spade, 1999] Spade, P. V. (Ed.). (1999). *The Cambridge companion to Ockham*. Cambridge University Press. DOI: 10.1017/CCOL0521583446

[Strachan et al., 2024] Strachan, J. W. A., Albergo, D., Borghini, G., et al. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7), 1285–1295. DOI: 10.1038/s41562-024-01882-z

[Taylor, 2014] Taylor, A. H. (2014). Corvid cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 5(3), 361–372. DOI: 10.1002/wcs.1286

[Templeton et al., 2024] Templeton, A., Conerly, T., Marcus, J., et al. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*. https://transformer-circuits.pub/2024/scaling-monosemanticity/

[Trinh et al., 2024] Trinh, T. H., Wu, Y., Le, Q. V., He, H., & Luong, T. (2024). Solving olympiad geometry without human demonstrations. *Nature*, 625(7995), 476–482. DOI: 10.1038/s41586-023-06747-5

[Turing, 1950] Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460. DOI: 10.1093/mind/LIX.236.433

[Wang et al., 2022] Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2022). Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. *arXiv preprint arXiv:2211.00593*. DOI: 10.48550/arXiv.2211.00593