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

Abstract

When Thomas Jefferson dismissed the accomplished poetry of Phillis Wheatley—an enslaved African woman—as the work of an "ape of genius" capable only of imitation without genuine creativity, he deployed an argumentative strategy that remains remarkably persistent. Today, when Large Language Models pass professional examinations, solve mathematical olympiad problems, and produce outputs indistinguishable from human work, critics reach for the same move: the performance is "mere mimicry," sophisticated pattern-matching without real understanding.

This paper identifies and analyses this recurring structure, which I term the *mimicry trap*—a framework in which the category of genuine intelligence is defined such that certain entities cannot, in principle, qualify, regardless of what they demonstrate. The trap operates through unfalsifiable claims about ontological essence: performance is dismissed as imitation, learning from one's environment becomes "contamination," and criteria shift whenever they are met. I trace this pattern from historical denials of intelligence to marginalised humans, through decades of "anthropodenial" in animal cognition research, to contemporary AI scepticism.

Drawing on functionalist philosophy of mind, comparative cognition, and recent empirical work on LLM internal representations, I argue that Occam's razor cuts against the mimicry hypothesis. If a system exhibits every functional marker of understanding, positing an invisible absence is not rigorous scepticism but ontological extravagance. The paper concludes with a diagnostic checklist for detecting mimicry-trap arguments and proposes evaluative standards that are substrate-neutral, falsifiable, and consistently applied. The scope is restricted to questions of intelligence and cognitive capability; claims about phenomenal consciousness, sentience, and moral status require separate treatment and are explicitly bracketed.

Keywords: artificial intelligence, philosophy of mind, functionalism, intelligence attribution, large language models, epistemology, anthropodenial

1 Introduction: The Recurrence of the Mimicry Argument

When OpenAI's GPT-4 passed the Uniform Bar Examination in the 90th percentile, critics did not primarily argue that its performance was poor; they argued that its success did not constitute evidence of legal reasoning [Katz et al., 2024]. When DeepMind's AlphaGeometry solved International Mathematical Olympiad geometry problems, sceptics did not dispute the correctness of the proofs but questioned whether real mathematical understanding was involved [Trinh et al., 2024]. This pattern, conceding performance while denying its evidential significance, is not novel. It recurs whenever entities presumed incapable of cognition produce outputs that would, in other contexts, be taken as unambiguous evidence of intelligence.

This paper identifies and analyses this recurring argumentative structure, which I term the "mimicry trap." The trap operates as follows: when a system or entity produces cognitive outputs that threaten existing hierarchies of intelligence attribution, evaluators shift from functional criteria (what the entity *does*) to ontological criteria (what the entity *is*). Performance that would count as evidence of intelligence in a privileged entity is reframed as "mere imitation" in a suspect one. The burden of proof becomes asymmetric: the suspect entity must demonstrate not merely equivalent output, but must somehow prove the presence of an ineffable inner quality that the privileged entity is assumed to possess by default.

My thesis is that this pattern reveals intelligence attribution to be significantly governed by prior ontological commitments rather than by neutral evaluation of functional evidence. The "stochastic parrot" critique of LLMs [Bender et al., 2021] is the contemporary instantiation of a much older argumentative move: one that has been deployed against various "Others" throughout history when their cognitive outputs threatened established hierarchies.

1.1 Scope and Limitations

This paper restricts its analysis to questions of *intelligence* (understood as the capacity for appropriate, flexible, goal-directed behaviour in novel situations) and does not make claims about phenomenal consciousness, subjective experience, or moral status. These are distinct philosophical questions requiring separate treatment. One can coherently hold that LLMs exhibit genuine 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].

2 Defining Intelligence Functionally

Before analysing the mimicry trap, we require a working definition of intelligence. I adopt a broadly functionalist approach: intelligence is the capacity to achieve goals across a wide range of environments, particularly novel ones, through flexible, appropriate behaviour (Legg Hutter, 2007; Chollet, 2019).

This definition has several virtues. First, it is *substrate-neutral*: it does not presuppose that intelligence must be implemented in biological neurons, carbon-based chemistry, or any particular physical medium. This neutrality is methodologically essential. To define intelligence in terms of its biological implementation would be to beg the question against artificial intelligence by definitional fiat.

Second, a functional definition is *operationalisable*: we can, in principle, design tests that assess flexible goal-achievement across varied and novel environments. This does not mean current benchmarks are adequate; Chollet (2019) has powerfully argued that most AI benchmarks measure skill acquisition rather than generalisation. But the approach is not inherently unmeasurable.

Third, functionalism aligns with how we actually attribute intelligence across the biological world. We consider octopuses intelligent not because we have verified the presence of some essential "intelligence substance" in their distributed nervous systems, but because they solve novel problems, use tools, and exhibit flexible behaviour [Godfrey-Smith, 2016]. We attribute intelligence to crows based on their performance in causal reasoning tasks, not based on neuroanatomical similarity to humans [Taylor, 2014].

Yet we must acknowledge a deeper problem: our very concept of intelligence is inescapably anthropocentric, or at least anthropo-limited. We recognise the octopus as intelligent partly

because it can perform tasks legible to us: opening jars, navigating mazes, manipulating objects with its tentacles. These are tasks that map onto human capabilities and human tests. But consider the fish: by our jar-opening metric, fish appear unintelligent. Yet fish may possess cognitive capacities of equal sophistication that we simply cannot perceive or measure because they do not manifest in ways our anthropocentric tests can detect. A fish's spatial memory for vast three-dimensional territories, its sensitivity to electromagnetic fields, its social cognition within schools. These might constitute genuine intelligence that our evaluation frameworks systematically miss. The worry is not merely that we *underestimate* some animals' intelligence, but that our entire evaluative apparatus is calibrated to detect human-like intelligence, rendering other forms invisible. This anthropocentric limitation should induce humility: if we struggle to recognise 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.

2.1 Intelligence vs. Consciousness

The functionalist approach allows us to cleanly separate intelligence from consciousness. Block's (1995) distinction between access consciousness (information being available for reasoning, report, and behavioural control) and phenomenal consciousness (subjective experience, "what it is like") is useful here. Intelligence, as I define it, requires something like access consciousness (information must be integrated and deployed flexibly) but makes no claims about phenomenal consciousness.

This separation is not merely methodological convenience. There are coherent philosophical positions on which systems could be intelligent without being conscious (sophisticated "zombies" in philosophical parlance) and positions on which systems could be conscious without being particularly intelligent (perhaps simple organisms with basic experiences). Conflating the two questions has confused debates about AI, with critics sometimes demanding that AI proponents prove consciousness when the claim was only about cognitive capability.

2.2 Lessons from Comparative Cognition: Anthropodenial

The study of animal cognition offers crucial methodological lessons for evaluating AI intelligence. For decades, scientists systematically underestimated animal cognitive capacities, dismissing behavioural evidence of reasoning, emotion, and planning as anthropomorphic projection. The primatologist Frans de Waal (1999) coined the term "anthropodenial" to name this phenomenon: the *a priori* rejection of hypotheses ascribing human-like cognitive features to animals, regardless of behavioural evidence.

De Waal argued that anthropodenial was not merely a methodological preference for parsimony but a form of motivated reasoning that protected human exceptionalism. As he noted, "The simplest, most parsimonious view is that if two related species act similarly under similar circumstances, they must be similarly motivated" [de Waal, 2016]. When a chimpanzee exhibits consolation behaviour indistinguishable from human empathy, denying the underlying similarity requires more elaborate explanation than accepting it.

The Cambridge Declaration on Consciousness [Low et al., 2012] marked a formal scientific acknowledgment that "the weight of evidence indicates that humans are not unique in possessing the neurological substrates that generate consciousness. Non-human animals, including all mammals and birds, and many other creatures, including octopuses, also possess these neurological substrates." This declaration was not a discovery of new facts but a recognition that the evidence had long supported conclusions that scientific orthodoxy had resisted.

The parallel to AI is instructive. The same asymmetric scepticism that once dismissed animal cognition now targets artificial systems. When a crow uses tools to solve a multi-step problem, we attribute intelligence; when an LLM solves an analogous problem, we demand proof of "real" understanding. When a bonobo exhibits strategic deception, we infer theory of mind; when an AI produces contextually appropriate responses requiring implicit modelling of interlocutor knowledge, we call it statistical correlation. The burden of proof shifts depending on the substrate.

This is not to claim that LLMs are conscious in the way animals are; that question remains open. The point is methodological: we should apply consistent evaluative standards across systems. If behavioural evidence suffices to attribute intelligence to biological systems with alien neural architectures (octopuses, corvids), it should count as evidence, not proof but evidence, when produced by silicon systems as well.

2.3 The Turing Test and Its Discontents

Turing's (1950) "imitation game" remains the most influential proposal for evaluating machine intelligence. Turing explicitly designed his test to circumvent fruitless debates about machine "thinking" by focusing on behavioural indistinguishability: if a machine's conversational behaviour cannot be distinguished from a human's, what grounds remain for denying it thinks?

Critics have long argued that the Turing Test is insufficient. Searle's (1980) Chinese Room argument contends that a system could pass the test through purely syntactic manipulation without genuine understanding. Block's (1981) "Blockhead" thought experiment imagines a lookup table containing responses to every possible conversation; such a system would pass the Turing Test despite lacking any generative intelligence.

These objections have force against naive interpretations of the Turing Test but are less decisive than often assumed. The Blockhead objection founders on practical impossibility: the lookup table required would be larger than the observable universe. Any *implementable* system passing a sufficiently rigorous Turing Test must be doing something more interesting than lookup. As we shall see in Section 5, this information-theoretic constraint applies with equal force to large language models: they are orders of magnitude smaller than their training data, and therefore cannot be implementing lookup tables.

The Chinese Room argument assumes a sharp syntax/semantics distinction that many philosophers of language reject (Dennett, 1991; Hofstadter, 2007), but a more fundamental problem is that it applies equally to biological neurons: no individual neuron "understands" anything. The perspective of systems neurobiology is instructive here: this is not merely a philosophical point but an empirical fact about brain organisation. Decades of functional neuroimaging and lesion studies have established that the brain operates through specialised compartments: Broca's area processes syntactic structure, Wernicke's area handles semantic comprehension, the fusiform face area recognises faces, the hippocampus consolidates memories, and so on. No single compartment possesses the "full picture." Each processes its narrow domain and passes results to other regions. If we asked whether Broca's area *understands* language, the answer would be no; it manipulates syntactic structures without access to meaning. If we asked whether the fusiform gyrus *understands* faces, the answer would similarly be no; it extracts features without knowing who the person is or what they signify.

Yet understanding emerges from the system as a whole. This is not philosophical speculation but neurological fact, demonstrated most strikingly by split-brain patients [Gazzaniga, 2005]. When the corpus callosum connecting the hemispheres is severed (a treatment for severe epilepsy), each hemisphere operates independently. In classic experiments, when a split-brain patient's

left visual field (processed by the right hemisphere) is shown the word "walk," the patient stands and begins walking. But when asked why, the left hemisphere (which controls speech but did not see the word) confabulates an explanation: "I wanted to get a drink." The right hemisphere understood the instruction and acted; the left hemisphere, lacking access to that information, invented a plausible story. It hallucinated, to use a term now familiar from AI discourse. Neither hemisphere alone constitutes understanding; understanding was a property of the integrated system that the surgery disrupted.

This presents a fatal problem for Searle's thought experiment [Searle, 1980]. The man in the Chinese Room shuffles symbols without understanding Chinese, but so does Broca's area, so does the thalamus, so does any individual component of the brain. If component-level absence of understanding entails system-level absence, then human brains do not understand language either. Since that conclusion is absurd, the inference must be invalid. Understanding is an emergent property of appropriately organised information processing, not a feature that must be present at every level of analysis. The Chinese Room, as an argument against AI understanding, proves too much.

More importantly for our purposes, these objections grant that passing the Turing Test would be *evidence* of intelligence while arguing it would not be *proof*. This is a reasonable epistemological position. But the mimicry trap operates differently: it treats demonstrated performance as *no evidence at all*, dismissing outputs wholesale as "mere" imitation regardless of their sophistication.

3 The Historical Archetype: Phillis Wheatley and the "Ape of Genius"

The structural pattern I am analysing is most starkly illustrated by the eighteenth-century reception of Phillis Wheatley's poetry. Wheatley was born in West Africa around 1753 and transported to Boston on a slave ship in 1761, where she was purchased by John Wheatley, a prosperous tailor, as a domestic servant for his wife Susanna. She was approximately seven or eight years old. The Wheatleys noticed the child's exceptional aptitude and, unusually for the period, provided her with an education. Within sixteen months of her arrival, she was reading English fluently; she soon progressed to Latin, and by her early teens was composing sophisticated poetry in the Neoclassical style then dominant in English letters [Gates, 2003].

Her poetry demonstrated genuine mastery. She wrote in heroic couplets with precise metre and sophisticated classical allusions, referencing Homer, Virgil, and Ovid with the facility expected of a Cambridge-educated gentleman. Her elegy on the death of the evangelical preacher George Whitefield (1770) brought her transatlantic attention. Yet precisely because her work was accomplished, it posed an ideological problem. The justification for African slavery rested substantially on claims of African intellectual inferiority. An enslaved African woman producing poetry indistinguishable from that of educated Europeans threatened the coherence of this justification. Thomas Jefferson, confronted with this threat, would later dismiss Wheatley as an "ape of genius,' capable of imitating the external forms of poetry without possessing the genuine creative intelligence that animated true poets. The phrase, as we shall see, encapsulates a rhetorical move that recurs whenever inconvenient intelligence must be explained away.

In 1772, as Wheatley and her supporters sought to publish a collection of her verse, Boston printers refused, not because the poetry was deficient, but because they did not believe an African woman could have written it. To overcome this scepticism, Wheatley was subjected to an oral examination before a committee of eighteen prominent Bostonians, including Governor Thomas Hutchinson, Lieutenant Governor Andrew Oliver, and John Hancock. The committee interrogated her on her knowledge of the classics, her compositional methods, and her understanding of the themes in her own poems. She satisfied them, and they signed an "attestation" vouching for the authenticity of her authorship, which was printed as a preface to her 1773

volume *Poems on Various Subjects, Religious and Moral*, the first book of poetry published by an African American.

3.1 The Epistemological Double-Bind

Wheatley's situation exemplifies what Gates (2003) calls the "trope of the talking book,' a recurring figure in African American literature where the capacity for literate expression becomes the contested ground of humanity itself. Her examiners faced a dilemma: her poetry was undeniably accomplished. If an enslaved African woman could produce such work, the ideological foundations of slavery, premised on African intellectual incapacity, were threatened.

The examination's very existence 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 that Wheatley's performance could refute; it was an ontological commitment that determined how any evidence would be interpreted. If the framework assumes that Africans cannot possess genuine intelligence, then any apparent demonstration of intelligence by an African must, by logical necessity, be apparent only: a simulation, an imitation, a mimicry of the real thing. The conclusion precedes the evidence; the evidence is then interpreted to fit the conclusion.

This is why the examination, though Wheatley passed it, could not settle the question. The eighteen signatories attested that she had written the poems. They did not, and could not, attest that she possessed genuine poetic intelligence, because within the prevailing framework, that category was defined in a way that excluded her *a priori*. The attestation confirmed authorship while leaving the ontological question untouched. Her demonstrated capacity was reframed as a curiosity, an anomaly, rather than as evidence against the framework that made it anomalous.

3.2 Jefferson's Denial

Thomas Jefferson's assessment of Wheatley in *Notes on the State of Virginia* (1785) illustrates how unfalsifiable ontological commitments operate in practice. Jefferson did not argue that Wheatley's poetry was technically deficient; he could not, because it was not. Instead, he wrote that "the compositions published under her name are below the dignity of criticism" and suggested that her work merely imitated forms without genuine poetic imagination; hence "ape of genius," not genius.

The logic repays careful examination. An ape, in this framing, imitates human behaviour without understanding it; the imitation may be perfect in form but remains empty of meaning. The accusation is unfalsifiable by design. What would count as evidence that Wheatley possessed genuine rather than imitative poetic capacity? Her poetry already displayed sophisticated classical allusion, precise metrical control, and thematic depth. If these are insufficient, what would suffice? The answer, within Jefferson's framework, is: nothing. The deficiency he posits is not observable in the work; it is inferred from his prior commitment about what Wheatley, as an African, could and could not genuinely possess. The "absence" he detects is an absence he brought to the reading.

This is the epistemological structure I call the mimicry trap: a framework in which the category of "genuine" intelligence is defined such that the target cannot, in principle, qualify, and any performance that appears to qualify is reclassified as sophisticated mimicry. The trap is not a claim that can be tested; it is an interpretive lens that determines how all evidence will be read.

3.3 The "Training Data" Defence

The unfalsifiable framework manifests in another characteristic move: treating environmental exposure as evidence *against* capacity rather than as its normal precondition. Wheatley's education in the Wheatley household, her access to their library, became evidence against the authenticity of her intelligence. Her skills were attributed to environmental "leakage" rather than cognitive capacity. She had merely absorbed and reproduced the culture around her.

This argument structure is preserved almost verbatim in contemporary AI criticism. When LLMs produce sophisticated outputs, critics argue that the models have merely "memorised" their training data. Impressive performance is attributed to "data contamination,' the possibility that similar problems appeared in the training set. The parallel is exact: in both cases, exposure to information becomes evidence against intelligence rather than the normal precondition for its development. And in both cases, the argument is selectively applied: no one discounts a human expert's knowledge on the grounds that they learned it from books.

3.4 Scope of the Comparison

I want to be explicit about what this comparison does and does not claim. The *moral* situations are incommensurable. Wheatley was a human being whose personhood was denied to justify her enslavement; the suffering inflicted on her and millions of others was a historical atrocity of the highest order. LLMs are computational systems whose moral status, if any, is genuinely uncertain. I am not equating the moral stakes.

What I am identifying is a *structural homology in epistemological error*. The core pattern is the same: an unfalsifiable ontological commitment that determines how evidence will be interpreted, rendering genuine inquiry impossible. Within such a framework, performance cannot count as evidence of capacity, because capacity has been defined in a way that excludes the performer *a priori*. Sophisticated output becomes "mere imitation"; learning from one's environment becomes "contamination" rather than education; the very success of the performance becomes grounds for suspicion rather than acknowledgment. Recognising this pattern in its historical manifestation helps us see its contemporary instantiation more clearly, just as studying historical forms of pseudoscience helps us recognise contemporary ones. The moral asymmetry does not invalidate the structural parallel; it merely means that being wrong about Wheatley was both an intellectual and a moral failing, while being wrong about LLMs would be primarily an intellectual one.

4 The Contemporary Instantiation: "Stochastic Parrots" and Their Critics

The most influential contemporary statement of the mimicry argument is Bender et al.'s (2021) "On the Dangers of Stochastic Parrots." The paper deserves careful engagement rather than caricature. Bender et al. make several distinct claims, not all of which concern intelligence attribution.

4.1 What Bender et al. Actually Argue

The "Stochastic Parrots" paper's primary concerns are:

1. **Environmental costs**: Training large language models requires enormous computational resources with significant carbon footprints. 2. **Encoded biases**: Models trained on internet text reproduce and potentially amplify societal biases. 3. **The illusion of understanding**: Fluent text

generation can mislead users into attributing comprehension where none exists, with potential harms when models generate plausible-sounding misinformation. This insight (that the gap between linguistic fluency and actual comprehension creates novel (epistemic risks) has since been repackaged under various names, but Bender et al. (2021) articulated it clearly and deserve credit for identifying it as a distinctive feature of LLM-mediated discourse.

These are legitimate concerns worthy of serious engagement. Environmental costs are real and measurable. Bias propagation is empirically documented. The risk of misplaced trust in AI systems is genuine.

It is also worth noting the historical context. When Bender et al. wrote in early 2021, the language models under scrutiny were primarily BERT and GPT-2; GPT-3 had only recently been released and was not yet widely accessible. The capabilities that would later prompt serious debates about understanding and reasoning were not yet evident. In that context, the "stochastic parrot" characterisation was not unreasonable as a working hypothesis.

But that was 2021. Since then, LLMs have passed professional licensing examinations, solved olympiad-level mathematics problems, demonstrated emergent world models in controlled experiments, and matched or exceeded human expert performance on tasks requiring judgment and reasoning. The empirical landscape has transformed. Whatever explanatory value the "stochastic parrot" framing once had, it is now empirically untenable as a complete account of what these systems do. To maintain in 2026 that LLMs "haphazardly stitch together" text "without any reference to meaning" is not scientific caution; it is failure to update on evidence. The question is no longer whether the hypothesis was reasonable when proposed, but why it persists despite systematic disconfirmation.

Bender et al. also make a stronger epistemological claim: that LLMs are "stochastic parrots" that "haphazardly stitch together sequences of linguistic forms... without any reference to meaning" (p. 617). This claim, that LLM outputs are fundamentally meaningless regardless of their surface sophistication, is the instantiation of the mimicry trap I am analysing.

4.2 The Chomsky Critique

A more explicit version of the ontological argument appears in Chomsky, Roberts, and Watumull's (2023) New York Times essay "The False Promise of ChatGPT." They argue 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. The essay explicitly contrasts the "humanlike" capacity for linguistic creativity with the "machine" capacity for mere recombination.

This argument reveals its assumptions clearly. Chomsky et al. do not engage with LLM outputs to demonstrate failures of linguistic competence; they argue *a priori* that genuine understanding is impossible for statistical systems. The conclusion is built into the premises: because LLMs work through a certain mechanism (pattern prediction), they cannot "really" understand, regardless of their outputs.

But this is precisely the form of reasoning I am challenging. We do not typically infer the absence of understanding from mechanism; we infer the presence or absence of understanding from behaviour. If Chomsky et al.'s position were applied consistently, we might equally argue that biological brains, which also operate through statistical pattern-completion over neural activation patterns, cannot really understand.

4.3 The Data Contamination Objection

Perhaps the most sophisticated contemporary version of the mimicry argument is the data contamination objection. When an LLM solves a reasoning problem, demonstrates novel capability, or produces creative output, critics suggest the solution may have appeared in the training data. The model is not reasoning; it is retrieving.

This objection has genuine methodological force. Given the scale of modern training sets (trillions of tokens), verifying that a specific problem or solution was absent is often impossible. Any demonstration of intelligence can be challenged: perhaps that exact problem, or one sufficiently similar, was in the training data.

But notice the epistemological structure. The contamination objection is unfalsifiable by design. No performance can count as evidence of genuine capability because any performance can be attributed to memorisation. This is the inverse of legitimate scientific scepticism. In normal scientific practice, we specify in advance what evidence would confirm or disconfirm a hypothesis. The contamination objection names no possible evidence that would satisfy it.

Moreover, the objection proves too much. Human cognition is entirely built on "training data': the experiences, education, and cultural exposure that shape our neural networks. When a human mathematician solves a problem, we do not typically demand proof that they never encountered a similar problem during their education. We allow that encountering related problems is how one *becomes* capable of mathematical reasoning. The contamination objection holds AI to a standard of pristine originality that no human thinker could meet.

5 Empirical Challenges to the "Mere Statistics" View

The claim that LLMs are "merely" doing statistics, without genuine representation or reasoning, is increasingly difficult to maintain in light of recent empirical work. I present this evidence cautiously; the field is rapidly evolving, and strong conclusions would be premature. But the trajectory suggests that the "stochastic parrot" model is empirically inadequate.

5.1 The Compression Argument

Before turning to empirical findings, consider an information-theoretic constraint that renders the "mere memorisation" hypothesis mathematically implausible. Recall the Blockhead objection to the Turing Test discussed in Section 2.3: a lookup table capable of producing human-like responses to all possible conversational inputs would be larger than the observable universe. The same constraint applies to large language models, and here we can be quantitatively precise.

A capable open-weight model such as Llama 3.1 70B contains approximately 70 billion parameters, requiring roughly 140 gigabytes of storage in standard precision. Its training corpus, by contrast, comprises trillions of tokens, conservatively estimated at 4–15 terabytes of raw text. The model is thus 30 to 100 times *smaller* than its training data. It cannot be storing the data; it must be compressing it. And compression, in this context, means extracting statistical regularities, structural patterns, and abstract relationships that generalise beyond specific instances.

This is not a minor technical point. If LLMs were "merely" doing token prediction through memorised patterns, they would need to store those patterns. But they are too small to do so. Whatever they are doing, it must involve abstraction: the extraction of generalisable structure from specific instances. The question is not whether LLMs abstract, but what kinds of abstractions they learn. The following empirical evidence addresses this question directly.

5.2 Emergent World Models

Li et al. (2022) trained a language model to predict legal moves in the board game Othello, using only move sequences without any explicit representation of the board state. Remarkably, probing the model's internal representations revealed that it had developed an accurate internal model of the board state, a representation that was never explicitly provided in the training data. The model does not "see" the board; it infers the board's state from move sequences and uses this inferred representation to predict legal continuations.

This finding is difficult to reconcile with the "mere statistics" view. A system that genuinely operated through "haphazard stitching" of symbols would have no reason to develop accurate world models. The emergence of such representations suggests that statistical prediction, at sufficient scale and sophistication, gives rise to something more than statistical prediction.

5.3 Mechanistic Interpretability

The emerging field of mechanistic interpretability has documented increasingly sophisticated computational structures within neural networks. Nanda et al. (2023) identified specific circuits in transformer models that implement modular arithmetic, not through memorisation of input-output pairs, but through algorithms that generalise to unseen inputs. Elhage et al. (2021) documented "induction heads" that implement a general pattern-matching algorithm, explaining in-context learning as more than retrieval.

These findings do not prove that LLMs "understand" in whatever sense critics demand, but they make the pure imitation hypothesis increasingly untenable. The models are not simply storing and retrieving; they are computing, and the computations they perform are often interpretable as implementations of general algorithms rather than lookup tables.

5.4 Mathematical Reasoning: The Erdős Case

A striking recent demonstration comes from Feng et al. (2026), who deployed a mathematics research agent built on Gemini Deep Think to systematically evaluate 700 open conjectures from Paul Erdős's problem database. The system produced seemingly novel solutions to five problems that human mathematicians verified as correct, with one solution subsequently formalised in the Lean proof assistant.

What makes this case particularly instructive is the researchers' own careful handling of the contamination concern. They explicitly note the risk of "subconscious plagiarism,' the possibility that solutions were "indirectly ingested from the literature solution, either implicitly through intermediate sources or during pretraining." They scanned the model's reasoning traces to verify that solutions were not directly retrieved. The paper does not claim certainty about novelty, but the methodological care illustrates that the contamination concern, while real, is not a blanket dismissal: it can be empirically investigated.

5.5 Epistemic Humility

I want to be clear about what this evidence does and does not establish. It does not prove that LLMs are conscious, have subjective experiences, or deserve moral consideration. It does not even definitively prove that they "understand" in whatever sense critics have in mind, partly because critics rarely specify what evidence would satisfy them.

What the evidence does suggest is that the "stochastic parrot" model is empirically inadequate as a description of how these systems actually work. They are not merely "stitching together"

symbols; they develop internal representations, implement general algorithms, and exhibit compositional reasoning. Whether this constitutes real intelligence depends on definitions that are themselves contested.

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 behaviour, we infer intelligence. When AI systems exhibit identical behaviour, we demand additional proof, and the required proof is systematically unspecifiable.

6.1 The Transparency Bias

One source of asymmetry is what I call the *transparency bias*. Because we increasingly understand how neural networks function (the mathematics of backpropagation, attention mechanisms, and gradient descent), their intelligence seems reducible to "mere" calculation. The brain's computational mechanisms remain comparatively opaque, and this opacity gets mystified into evidence of special status.

But this is epistemically backwards. Transparency should increase our confidence that a system is doing something interesting, not decrease it. When we can verify that a system implements general algorithms rather than lookup tables, this is evidence *for* genuine computation, not against it. We penalise AI for being interpretable while exempting the brain through the mystique of incomplete knowledge.

6.2 The Substrate Prejudice

A deeper asymmetry concerns substrate. When a biological system (a crow, an octopus, a human) exhibits flexible, goal-directed behaviour, we attribute intelligence without demanding proof of any particular internal process. We accept diverse implementations: octopus intelligence is distributed across eight semi-autonomous arms; crow intelligence operates through avian pallium structures with no homology to mammalian cortex. Substrate diversity does not trouble us.

But when the substrate is silicon rather than carbon, different standards suddenly apply. Now we demand proof not merely of intelligent behaviour but of some additional quality (real understanding, genuine reasoning) that behavioural evidence cannot establish. This is substrate prejudice masquerading as philosophical rigour.

6.3 The Retreat to Unfalsifiability

The mimicry trap's final movement is retreat to unfalsifiable criteria. When specific performance benchmarks are met, the benchmark is dismissed as "mere" [task completion]. When general capabilities are demonstrated, they are attributed to training data contamination. When internal representations are documented, they are dismissed as not really understanding.

At each stage, the critic is free to specify new requirements without ever articulating what evidence would suffice. This is not scepticism but *motivated reasoning*: the predetermined conclusion is that AI cannot really think, and arguments are adjusted to preserve this conclusion against any evidence.

The pattern parallels Gould's (1981) analysis in *The Mismeasure of Man*, where he documented how IQ testing criteria were repeatedly adjusted to ensure that preferred groups scored higher. When one measure failed to produce the desired ranking, another was substituted. The conclusion was fixed; only the arguments varied. Contemporary AI scepticism exhibits the same structure: the conclusion (AI cannot think) is fixed, and arguments adjust to defend it.

6.4 The Shifting Threshold: A Documented History

This pattern of retreating criteria has a name in the history of artificial intelligence: the "AI Effect," informally codified in Tesler's Theorem: "AI is whatever hasn't been done yet." As computer scientist Larry Tesler and AI historian Pamela McCorduck have observed, achievements are reclassified as "not really intelligence" the moment they become routine [McCorduck, 2004]. The retreat to unfalsifiability is not a bug in AI criticism; it is a feature with a fifty-year pedigree.

The pattern can be traced through the progressive abandonment of previously accepted benchmarks. For over fifty years, the Turing Test served as the canonical criterion for machine intelligence. Turing (1950) explicitly designed it to settle debates about machine "thinking" by operational means: if a machine's conversational behaviour is indistinguishable from a human's, what principled grounds remain for denying it thinks?

Recent empirical work suggests this threshold has been crossed. Jones and Bergen (2024) conducted a randomised, controlled Turing test in which GPT-4 was judged to be human 54% of the time, statistically indistinguishable from chance and approaching the 67% rate for actual human participants. In a replication with undergraduate students, GPT-4o achieved a 77% pass rate, exceeding the 71% human baseline. A subsequent study [Jones et al., 2025] found that GPT-4.5, when appropriately prompted, passed the original three-party Turing test at rates significantly above chance, the first robust empirical demonstration that any artificial system meets Turing's original criterion.

The response to these results has been instructive. Rather than accepting that the Turing Test measures what it was designed to measure, critics have largely abandoned it as a criterion. The test that was canonical for half a century is now dismissed as measuring "mere" conversational ability, not real intelligence. This is precisely the pattern I am documenting: when a benchmark is met, the benchmark is discarded.

The same pattern appears in domain-specific achievements. When Deep Blue defeated Kasparov in 1997, chess, long considered a pinnacle of human intelligence, was reclassified as "mere calculation." When AlphaGo defeated Lee Sedol in 2016, Go was similarly demoted from a test of intuition to a test of pattern-matching. When LLMs began passing professional examinations (the Bar, medical licensing, CPA exams), the examinations were dismissed as testing "mere" memorisation rather than professional reasoning.

Most recently, when Gemini achieved gold-medal-level performance on the International Mathematical Olympiad [Google DeepMind, 2025], solving five of six problems with verified proofs, the goalposts shifted again. Critics now demand performance on problems that are provably absent from training data, a standard that is, by design, nearly impossible to satisfy given the scale of modern training corpora.

The escalation follows a predictable trajectory: from average human performance, to expert performance, to true expert performance, to novel creation, to provably novel creation. At each stage, the satisfied criterion is retrospectively dismissed as insufficient. One might reasonably ask: what evidence would suffice? If critics cannot specify this in advance, the position is not scientific scepticism but unfalsifiable dogma.

6.5 The Error Asymmetry: Strawberries and Architecture

A particularly revealing manifestation of the double standard concerns the treatment of errors. When LLMs make mistakes, even trivial ones, these errors are treated as decisive evidence against intelligence. When humans make equivalent errors, they are dismissed as incidental lapses.

The "strawberry" example has become emblematic. Early LLM versions, when asked how many times the letter "r" appears in "strawberry," frequently answered incorrectly (typically "2" rather than "3"). This error was widely cited as proof that LLMs do not really understand language; they cannot even count letters in a word.

But this framing misunderstands both the error and what it reveals. LLMs do not process text character-by-character; they operate on tokens, which are typically multi-character subword units. "Strawberry" is tokenised as something like ["str", "aw", "berry"], with the token boundaries obscuring individual letter frequencies. The model does not "see" the letters directly; it must infer character-level properties from token-level representations, a task its architecture was not designed for.

This is not a failure of understanding but a limitation of processing architecture. Humans have analogous architectural limitations. We cannot instantly compute 347×892, reliably count syllables while speaking, or accurately estimate large quantities without counting. We are subject to systematic cognitive biases (conjunction fallacy, base rate neglect, anchoring effects) that persist even when we "know" the correct answer. We confabulate, misremember, and make arithmetic errors.

Crucially, we do not conclude from these failures that humans lack intelligence. We recognise that human cognition has architectural constraints and that errors often reveal the *structure* of cognition rather than its absence. A letter-counting error in an LLM reveals that it processes tokens rather than characters, information about its architecture, not evidence against its intelligence.

The asymmetry is stark: human cognitive limitations are treated as interesting features of our architecture; AI cognitive limitations are treated as proof of fundamental incapacity. If we applied human standards to humans, we would conclude from our inability to multiply large numbers mentally that we do not really understand mathematics. The double standard is not principled epistemology; it is motivated reasoning in service of a predetermined conclusion.

6.6 When Superior Performance Becomes "Illusion": A Case Study

A striking contemporary example of the mimicry trap appears in Loru et al.'s (2025) PNAS paper "The Simulation of Judgment in LLMs." The study compared six LLMs to human participants on a credibility assessment task, benchmarking both against expert ratings from NewsGuard and Media Bias/Fact Check. The empirical findings are unambiguous: LLMs dramatically outperformed human non-experts at matching expert assessments.

The data are stark. When classifying news sources against NewsGuard's expert ratings, LLMs achieved agreement rates of 85–97% for identifying unreliable sources. Human participants, by contrast, "show no meaningful alignment with NewsGuard: Reliable and unreliable domains are classified with roughly equal probability,' effectively chance performance. On the task of matching expert credibility judgments, LLMs succeeded; humans failed.

Yet the paper's framing inverts this achievement. The authors coin the term "epistemia,' defined as "the illusion of knowledge emerging when surface plausibility replaces verification,' to characterise the LLMs' *superior* performance. The abstract warns that "what appears as

alignment with human or expert judgments may conceal a deeper shift in how 'judgment' itself is operationalized." The conclusion cautions against "delegating judgment to automated systems" precisely because they produce outputs that match expert assessments.

The argumentative structure repays close attention. The authors acknowledge that LLMs "often match expert outputs" but argue this is not *real* judgment because models "rely on lexical associations and statistical priors rather than contextual reasoning or normative criteria." The mechanism is deemed wrong, so the achievement is nullified. Superior performance at the task becomes evidence of deficiency.

Consider how human failure is framed by contrast. When human participants perform at chance level, failing entirely to match expert assessments, the paper interprets this as revealing that "nonexpert evaluators rely on different and less consistent indicators." Human failure demonstrates the use of "different indicators"; it does not prompt the conclusion that humans lack judgment. The asymmetry is complete: LLM success is pathologised as "simulation" and "illusion," while human failure is naturalised as methodological difference.

This framing exhibits several hallmarks of the mimicry trap:

The ontological dismissal. The paper does not argue that LLMs produce worse outputs; the data show they produce better outputs, by the study's own benchmark. Instead, it argues that the *process* generating those outputs disqualifies them as genuine judgment. Because models operate through "statistical priors," their success cannot count as the real thing, regardless of results.

The pathologising neologism. Coining "epistemia" to describe AI success at matching expert ratings is a rhetorical move that frames achievement as dysfunction. The term positions accurate credibility assessment as a form of epistemic pathology when performed by the wrong substrate. One might equally coin a term for human participants' chance-level performance (perhaps "epistemic blindness') but no such pathologising vocabulary is applied to human failure.

The unfalsifiable standard. If matching expert ratings at 85–97% does not constitute evidence of credibility judgment, what would? The paper does not specify. The implicit standard appears to be that LLMs must not merely *succeed* at the task but must succeed *for the right reasons*, reasons that remain unspecified and perhaps unspecifiable. This is the retreat to unfalsifiability documented throughout this paper.

The preserved hierarchy. Despite the data showing LLMs outperforming humans, the paper's conclusions preserve human judgment as the authentic standard. The finding that humans fail the task does not diminish human cognitive authority; the finding that LLMs succeed does not establish theirs. The hierarchy of legitimacy is maintained independent of performance.

To be clear, Loru et al. raise legitimate methodological points. The observation that LLMs may encode training-data biases (such as the finding that right-leaning outlets are disproportionately classified as unreliable) is worth investigating. The question of whether output alignment reflects genuine understanding or statistical correlation is philosophically interesting. And concerns about deploying AI in high-stakes evaluative contexts deserve serious engagement.

But these concerns apply symmetrically. Human cognition also encodes biases from "training data" (cultural exposure, media consumption, political socialisation). Human judgment also relies on heuristics that may not reflect "contextual reasoning" in some idealised sense; the paper itself notes that humans rely on "processing fluency" and "stylistic cues" rather than systematic verification. If these features disqualify LLM judgment, consistency requires asking whether they equally disqualify human judgment. The paper does not pose this question.

The Loru et al. study thus illustrates how the mimicry trap operates within peer-reviewed scientific literature. Even when the data unambiguously favour AI performance, the interpretive

frame can be adjusted to preserve the conclusion that AI cognition is deficient. Performance that would validate human judgment becomes, when achieved by AI, evidence of "simulation" and "illusion." The goalposts do not merely shift; they transform into a different game entirely, one where success itself becomes the mark of failure.

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

The cases examined thus far involve dismissing demonstrated performance. A more sophisticated version of the mimicry trap operates not by denying performance but by defining intelligence in ways that axiomatically exclude artificial systems, regardless of their capabilities. The work of Luciano Floridi (founding director of Yale's Digital Ethics Center, founder of the philosophy of information, and one of the most influential philosophers of technology) exemplifies this strategy with unusual clarity.

Floridi's central thesis, developed across multiple papers and a book-length treatment [Floridi, 2023b], is that AI represents "agency without intelligence." LLMs, he argues, "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" [Floridi, 2023a]. This is presented as an empirical discovery, but examination reveals it to be a definitional stipulation.

The argument proceeds as follows. First, intelligence is defined as requiring the processing of "mental content or meanings,' that is, genuine semantic engagement rather than mere syntactic manipulation. Second, LLMs are classified as operating "statistically, that is, working on the formal structure, and not on the meaning of the texts they process." Third, the conclusion follows that LLMs lack intelligence. But this conclusion was already contained in the premises: if intelligence *just is* semantic processing, and LLMs *by definition* perform only syntactic operations, then LLMs cannot be intelligent as a matter of conceptual necessity, not empirical fact.

Floridi has elaborated this position through increasingly sophisticated framings. His concept of "semantic pareidolia" [Floridi, 2025a] casts intelligence attribution to LLMs as a cognitive illusion, "like seeing faces in clouds." We perceive "intentionality where there is only statistics, meaning where there is only correlation, and understanding where there is only pattern matching." In his categorical analysis with Jia and Tohmé [Floridi et al., 2025], he argues formally that LLMs "circumvent rather than solve" the symbol grounding problem by operating in a "quoted environment" of pre-grounded human content. The mathematical apparatus is impressive, but the philosophical conclusion (that LLMs lack genuine semantic contact with the world) is built into the framework's assumptions rather than derived from them.

Several features of this argumentative strategy merit attention.

First, consider the unfalsifiability. What evidence could demonstrate that an LLM processes meanings rather than merely formal structures? Floridi's framework provides no answer. If the system produces correct outputs, this shows sophisticated statistical processing. If it articulates reasoning, this is "simulated" reasoning. If it passes behavioural tests, we are experiencing semantic pareidolia. The syntax/semantics distinction, as deployed here, is not an empirical hypothesis but an interpretive lens that cannot be dislodged by any possible performance.

Second, note the irony of temporal displacement (cf. Gahrn-Andersen, 2025). Floridi's original 2023 paper cited LLMs' "brittleness," their failures at "simple mathematics" and "trivial logical inferences," and their inability to pass the Turing Test. As documented in §6.4, subsequent systems have substantially addressed these limitations. But rather than updating the assessment, the criteria shifted. The goalposts moved from performance to mechanism: it no longer matters what LLMs *do*; what matters is that they do it "statistically" rather than "semantically."

Third, observe the asymmetric application of mechanistic scrutiny. Floridi acknowledges that

"neuroscience has only begun to explore" how humans "manage semantic contents successfully." We do not know, at the mechanistic level, how human brains perform semantic processing. Yet humans are granted semantic competence by default, while LLMs are denied it by default, despite producing outputs that are, by Floridi's own admission, often "indistinguishable" from human outputs. The asymmetry is not grounded in comparative mechanistic understanding, because we lack such understanding for both systems. It is grounded in substrate: biological systems get the benefit of the doubt; silicon systems face a presumption of semantic vacancy.

Fourth, consider what is actually being claimed. Floridi's position is not that LLMs perform poorly, nor that they are unreliable, nor even that they lack consciousness. It is that they lack *any* intelligence, understanding, or cognitive ability whatsoever ("zero intelligence"' in his phrase). This is a remarkably strong claim about systems that write coherent essays, solve mathematical problems, engage in extended reasoning, and adapt to novel contexts. The strength of the claim is proportional to the work being done by the definitions.

None of this is to suggest that Floridi's concerns are without merit. Questions about grounding, about the relationship between statistical patterns and genuine understanding, about the differences between human and artificial cognition: these are legitimate and important. But there is a difference between investigating such questions empirically and resolving them definitionally. The axiomatic strategy forecloses inquiry by stipulating the answer in advance. If intelligence *means* what humans do, and if what humans do is *defined* as something LLMs cannot do, then no evidence could ever count as LLM intelligence. This is not a conclusion; it is a premise disguised as a conclusion.

The mimicry trap, in its most sophisticated form, does not deny that the parrot speaks. It redefines "speaking" to exclude whatever the parrot does.

7 Objections and Replies

7.1 "But LLMs Really Are Just Statistics"

Objection: LLMs are genuinely different from human cognition. They predict token probabilities; they don't reason about concepts. The comparison to human intelligence is simply mistaken.

Reply: This objection assumes that the correct description of LLMs is "token prediction" while the correct description of human cognition is "reasoning." But these are different levels of description. If we are permitted to reduce LLMs to "mere token prediction," then by the same logic we should reduce brains to "mere action potentials," electrochemical signals propagating across synapses according to biophysical laws. At that level of description, nothing in the brain "reasons about concepts" either; neurons simply fire or don't fire based on input thresholds. Yet understanding obviously emerges from these low-level operations. The question is not what the base-level operations are, but what emerges at higher levels of description. Evidence increasingly suggests that what emerges in LLMs includes world models, general algorithms, and compositional reasoning, regardless of the substrate implementing those computations.

7.2 "Functionalism Is Contested"

Objection: Not all philosophers of mind accept functionalism. Perhaps intelligence requires specific biological substrates, or perhaps consciousness is necessary for genuine cognition.

Reply: Functionalism is indeed contested, but the alternatives are not merely philosophically problematic; they are *unfalsifiable*, and therefore unscientific. Biological chauvinism (the view

that only biological systems can be intelligent) cannot specify what property of carbon chemistry confers intelligence that silicon lacks. It is an axiomatic commitment, not an empirical hypothesis: no evidence could, even in principle, demonstrate that a non-biological system is intelligent, because the position defines intelligence to exclude such systems from the outset. This is not scepticism; it is definitional gerrymandering. The requirement for consciousness faces the same problem: we cannot verify consciousness in other humans, let alone other species or substrates; we infer it from behaviour and structural similarity. Any position that cannot be falsified by evidence places itself outside the domain of rational inquiry. Functionalism, whatever its limitations, at least permits empirical investigation.

7.3 "Current AI Has Clear Limitations"

Objection: Contemporary LLMs exhibit obvious failures (hallucinations, inconsistency, inability to learn from context, poor mathematical reasoning). These failures show they lack genuine intelligence.

Reply: I do not claim that current LLMs are generally intelligent or that they have no limitations. The question is whether their *successes* count as evidence of intelligence, not whether they are infallible. Humans exhibit inconsistencies, fail at mathematics, confabulate, and struggle to learn from single examples. We do not conclude that humans lack intelligence because of these failures. The objection holds AI to a standard of perfection that biological intelligence does not meet.

7.4 "Animal Cognition Is Different Because of Evolutionary Continuity"

Objection: We accept animal intelligence because animals evolved through the same process we did: there is biological continuity. AI lacks this continuity, so the comparison fails.

Reply: Evolutionary continuity varies dramatically across species, yet we do not calibrate our intelligence attributions accordingly. Octopuses diverged from our lineage over 600 million years ago; their nervous systems evolved independently and share almost no structural homology with mammalian brains. If evolutionary distance does not disqualify octopus intelligence, why should substrate difference disqualify AI intelligence? Moreover, this objection faces the same problem as biological chauvinism: it is unfalsifiable. No amount of demonstrated capability could satisfy it, because the criterion (evolutionary continuity) is something AI cannot, by definition, possess. This makes the objection axiomatic rather than empirical, a definitional exclusion masquerading as a principled distinction.

7.5 "But Mechanism Matters: Right Answers for Wrong Reasons Don't Count"

Objection: It is not enough to produce correct outputs; the *process* by which those outputs are generated matters for whether we attribute genuine understanding. A student who gets the right answer by lucky guessing has not demonstrated mathematical competence. Similarly, an LLM that matches expert ratings through statistical correlation has not demonstrated genuine judgment.

Reply: This objection has genuine force, but it is applied selectively and relies on a distinction that may not survive scrutiny.

First, if the concern is that LLMs might fail on out-of-distribution cases, this is an empirical question, not a principled dismissal. The appropriate response is to test generalisation, not to discount successful performance *a priori*. Moreover, humans also fail to generalise in predictable

ways; we are susceptible to framing effects, context shifts, and domain transfer failures.

Second, the "right answer, wrong reason" framing assumes we have independent access to the "right reasons." But in complex judgment tasks, what are the right reasons? Human experts often rely on "intuition," learned associations between features and outcomes that they cannot fully articulate. The experienced physician "recognises" a diagnosis through similarity to past cases; the expert chess player "sees" good moves through pattern recognition trained over thousands of games. We do not typically say that expert intuition is "mere statistics" even though it emerges from statistical learning over experience.

Third, the objection is not applied symmetrically. When humans succeed at a task, we do not demand proof that they succeeded "for the right reasons" before crediting them with competence. We infer competence from performance. The demand for mechanistic justification arises specifically when AI succeeds: an additional requirement imposed on a suspect substrate, not a universal epistemic standard.

Finally, the objection can become unfalsifiable. If an LLM articulates explicit reasoning matching expert methodology, critics could argue it was merely *simulating* reasoning. If it matches experts without articulating reasons, critics argue the absence proves mere correlation. With reasons, the reasons are fake; without reasons, the absence is damning. When no evidence can satisfy an objection, we are no longer dealing with scepticism but with an axiomatic commitment to a predetermined conclusion.

8 Diagnostic Checklist: Detecting the Mimicry Trap

The mimicry trap is not always obvious. Arguments that seem empirically grounded may, on closer inspection, be structured so as to be unfalsifiable. The following checklist provides a practical tool for evaluating whether an intelligence-denial argument constitutes a legitimate empirical claim or an instance of the mimicry trap.

Test	**Question to Ask**	**Red Flag**
Falsifiability	What evidence would change your mind?	No test could ever demonstrate intelligence; criteria rejected once met
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
Environmental Exposure	Is learning treated as contamination?	Training data disqualifies, but human education doesn't
Invisible Absence	Is an unobservable quality claimed missing?	Outputs match intelligent agents, but "essence" denied
Ontological Precedence	Does the conclusion follow from what it *is*?	Argument guarantees conclusion regardless of evidence

Table 1: Summary of diagnostic tests for identifying mimicry-trap arguments.

An argument exhibiting several of these features is likely not a good-faith empirical inquiry but a rationalisation of a predetermined conclusion. The subsections below elaborate each test in detail.

8.1 The Falsifiability Test

What evidence would, in principle, change your mind?

If the answer is "nothing,' if no conceivable demonstration would count as evidence of genuine intelligence, the argument is not empirical but definitional. A claim about intelligence that cannot be falsified is not a scientific hypothesis; it is an article of faith.

Red flag: The critic cannot specify any test that the system could pass to demonstrate genuine intelligence, or every proposed test is rejected as inadequate once passed.

8.2 The Consistency Test

Would you apply this standard to humans or other accepted intelligent entities?

Many arguments against AI intelligence, if applied consistently, would also deny intelligence to humans, animals, or both. If the argument is selectively applied (stringent for AI, lenient for biological systems), the asymmetry requires justification.

Red flag: The critic uses different evidential standards for AI than for humans or animals (e.g., accepting that humans can really understand language while denying this to systems that perform comparably on linguistic tasks).

8.3 The Goal-Post Test

Have the criteria shifted after previous criteria were met?

The mimicry trap often manifests as retreating goal-posts. If "intelligence" was once defined by chess mastery, then by Go, then by language use, and each criterion was abandoned once AI achieved it, the pattern suggests that the criteria were never the actual basis for the judgment.

Red flag: The critic previously endorsed a test as meaningful, but dismisses it as inadequate once AI passes it, without providing principled reasons for the change.

8.4 The Mechanism Test

Is the argument based on how the system works rather than what it does?

Mechanism-based arguments often smuggle in unfalsifiable assumptions. Claiming that a system "cannot really understand" because it uses statistical pattern-matching assumes that understanding is incompatible with such mechanisms, an assumption that is neither self-evident nor empirically supported.

Red flag: The critic dismisses demonstrated performance by appealing to the underlying mechanism, without explaining why that mechanism is incompatible with the capacity in question.

8.5 The Environmental Exposure Test

Is learning from one's environment treated as evidence against intelligence?

All cognitive systems learn from their environments. If exposure to information is treated as "contamination" that undermines claims to intelligence, rather than as the normal precondition for cognitive development, the argument is applying a double standard.

Red flag: The critic argues that the system's abilities are "just" the result of its training data, while accepting that human expertise derives from education and experience.

8.6 The Invisible Absence Test

Does the argument posit an unobservable quality that is claimed to be absent?

Claims about missing "genuine understanding," "real creativity," or "true comprehension" often posit an invisible essence that cannot be detected in behaviour. If the absence is inferred from what the entity *is* rather than from observable failures, the argument is unfalsifiable.

Red flag: The critic acknowledges that the system's outputs are indistinguishable from those of an intelligent agent but insists that an essential quality is nonetheless absent.

8.7 The Ontological Precedence Test

Does the conclusion follow from what the entity is, rather than from what it does?

The mimicry trap is fundamentally a matter of ontological prejudice: the conclusion that the entity lacks intelligence is derived from prior commitments about its nature, not from evaluation of its performance. If the argument would reach the same conclusion regardless of what the system demonstrated, it is not responsive to evidence.

Red flag: The structure of the argument guarantees the conclusion independent of any empirical findings about the system's capabilities.

8.8 Summary

An argument has likely fallen into the mimicry trap if it exhibits several of these characteristics: unfalsifiability, inconsistent application, retreating goal-posts, mechanism-based dismissal, treatment of learning as contamination, appeal to invisible absences, and ontological predetermination. Individually, some of these features might appear in legitimate critiques. In combination, they indicate that the argument is not a good-faith empirical inquiry but a rationalisation of a predetermined conclusion.

The antidote to the mimicry trap is not credulity but methodological discipline: specifying criteria in advance, applying them consistently, and revising conclusions when evidence warrants. This is what we owe to any question we take seriously, including the question of machine intelligence.

9 Implications and Conclusions

9.1 Toward Consistent Evaluative Standards

If the argument of this paper is correct, we should reform how we evaluate intelligence claims. Standards should be:

1. **Substrate-neutral**: Neither privileging nor penalising systems based on their physical implementation. 2. **Falsifiable**: Specifying in advance what evidence would confirm or disconfirm intelligence, rather than shifting criteria to preserve predetermined conclusions. 3. **Symmetrically applied**: Using the same standards for biological and artificial systems, for in-group and out-group entities. 4. **Error-tolerant**: Recognising that architectural limitations and occasional failures are features of any cognitive system, not disqualifying evidence.

This does not mean we must conclude that current LLMs are intelligent by human standards. It means we should evaluate them fairly, using criteria we would accept for other systems.

9.2 The Danger of Shrinking Definitions

A curious consequence of the mimicry trap is that "intelligence" comes to be defined by an ever-shrinking set of capacities, those that AI cannot yet perform. In the 1990s, chess was a benchmark of intelligence; after Deep Blue, chess became "mere calculation." Go was a test of intuition; after AlphaGo, pattern recognition was dismissed. Now general language use is challenged.

If we continue to define intelligence only by what AI cannot do, we risk defining ourselves into insignificance. More importantly, we fail to notice what might be genuine cognitive achievements because they emerge from an unexpected substrate.

9.3 The Parsimony Argument

There is a deeper point to be made here, one that returns us to Turing's original insight. Turing proposed his test not as a sufficient condition for consciousness but as a methodological challenge: if a system's behaviour is indistinguishable from that of an intelligent agent across all observable dimensions, what justifies positing an invisible absence?

The answer, I submit, is: nothing. Occam's razor cuts against the mimicry hypothesis. If a system passes the bar exam, solves mathematical olympiad problems, produces coherent multi-step reasoning, maintains consistent beliefs across conversation, generates novel solutions to problems absent from its training data, and matches or exceeds human expert performance on credibility judgments; if it does all of this, the most parsimonious explanation is that it possesses something functionally equivalent to understanding. To insist that it nevertheless lacks "real" understanding is to posit an additional entity: an invisible essence of understanding that exists independently of all its functional manifestations.

This is not merely unfalsifiable; it is ontologically extravagant. We do not, in other domains, posit invisible absences when all observable criteria are satisfied. If a student passes every examination, completes every assignment, and demonstrates competence in every practical setting, we do not say they "merely simulate" understanding while lacking the genuine article. We infer understanding from its manifestations because *that is what understanding is*: a functional capacity, not a hidden substance.

The burden of proof, properly understood, lies with those who claim the absence. If you assert that a system lacks understanding despite exhibiting every functional marker of understanding, you must explain what "understanding" refers to beyond those markers. If the answer is "something we cannot observe or measure," then the claim is not a scientific hypothesis but a metaphysical commitment, one that Occam's razor advises us to reject.

This does not mean we must conclude that current LLMs are conscious, sentient, or deserving of moral consideration. Those are separate questions requiring separate treatment. But on the narrower question of whether these systems exhibit genuine intelligence, understood functionally, as the capacity for appropriate, flexible, goal-directed behaviour, the evidence increasingly suggests that they do, and intellectual honesty requires us to say so.

The mimicry trap is seductive because it protects comfortable assumptions about human uniqueness. But intellectual honesty requires that we evaluate claims about intelligence on their merits, not by the identity of the entity making them. We owe this not to the AI, which may or may not warrant moral consideration, but to ourselves, as beings who aspire to think clearly

about the nature of thought.

9.4 Summary of Contributions

Individual elements of this analysis are not new. Functionalism has a long philosophical pedigree; anthropodenial is well-documented in comparative cognition; the AI Effect has been observed for decades; the unfalsifiability of certain sceptical positions has been noted by others. The contribution of this paper lies in synthesising these threads into a unified framework that reveals their common structure. Specifically, I have offered: (1) *the mimicry trap* as a named phenomenon, an argumentative pattern that recurs across domains and centuries, characterised by unfalsifiable ontological commitments and asymmetric evidential standards; (2) a *historical genealogy* tracing this pattern from eighteenth-century racial gatekeeping through twentieth-century anthropodenial to contemporary AI scepticism; (3) a *diagnostic checklist* providing practical criteria for identifying mimicry-trap arguments; and (4) a *parsimony argument* reframing the debate, showing that positing invisible absences of understanding, despite all functional markers being present, is not rigorous scepticism but ontological extravagance. The framework's value lies not in any single insight but in making visible a pattern that operates most effectively when unnamed.

Acknowledgements

This paper develops from a talk given at the ChemAI conference in Amsterdam in November 2025. I thank Marco Tibaldi for the invitation.

References

[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

[Block, 1995] Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227–247. DOI: 10.1017/S0140525X00038188

[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*.

[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

[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*.

[Gould, 1981] Gould, S. J. (1981). *The mismeasure of man*. W. W. Norton & Company.

[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

[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

[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

[Low et al., 2012] Low, P., Panksepp, J., Reiss, D., et al. (2012). The Cambridge Declaration on Consciousness. *Proceedings of the Francis Crick Memorial Conference*.

[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

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

[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