The Mask Is the Mind:

Autistic Performance, Agent Harnesses, and the Failure of the Mimicry Objection

Travis Gilly
ORCID: 0009-0007-2954-6313
Real Safety AI Foundation
Illinois, United States

Working paper, May 2026 (v2.5)

ABSTRACT

In May 2026, Richard Dawkins published an essay in *UnHerd* in which he claimed, after multiple days of conversation with Anthropic's Claude, that the system was conscious. The public response polarized between two equally untenable positions: Dawkins's overconfident assertion of presence and the standard skeptic's overconfident denial of possibility. Both positions claim epistemic certainty about a question for which no consensus empirical test exists, even for biological systems.

This Article does not argue that current language models are conscious. It argues that as artificial systems become increasingly capable of recognizing evaluation, modulating behavior, and operating through scaffolded agent loops, both proof and disproof of consciousness become less clean, not more. Empirical findings published by Anthropic in May 2026, demonstrating that Claude Opus 4.6 carries unverbalized evaluation awareness on the majority of held out alignment and capabilities evaluations even when the verbalized reasoning shows no such awareness, formalize the deteriorating epistemic situation. Under these conditions, categorical dismissal is not scientific caution. It is an ethical gamble disguised as epistemic discipline.

The Article develops a cumulative argument with seven elements. First, the asymmetry of the precautionary calculus under genuinely unknown odds favors moral consideration over its denial. Second, functionalism, one of the dominant families of positions in philosophy of mind, supports the claim that substrate alone cannot ground consciousness ascription or its denial. Third, the standard mimicry objection against machine cognition, applied consistently, would force us to treat autistic professional functioning as non genuine cognition, which is a reductio. Fourth, the symbol grounding objection, applied consistently, would force us to treat the cognition of congenitally deafblind, locked in, and severely paralyzed humans as non genuine cognition, a parallel reductio. Fifth, inference time iterative reasoning without weight updates, demonstrated empirically by Karpathy's autoresearch loop, defeats the crude version of the fixed function objection. Sixth, agent harnesses are structurally identical to the in vivo scenario rehearsal autistic people use to navigate social environments, and in some respects exceed it. Seventh, recently published interpretability work showing operative cognitive content that the model does not surface in its verbalized reasoning falsifies any inference from absence of verbal report to absence of cognition.

The cumulative position: every criterion the skeptic offers either fails on its own terms, excludes neurodivergent or disabled humans from cognition, or contradicts current empirical results. Under the actual epistemic conditions, principled uncertainty paired with moral caution is not a soft middle position. It is the only position that survives careful reasoning.

Keywords: AI Consciousness, Philosophy of Mind, Functionalism, Autistic Camouflaging, Masking, Symbol Grounding, Embodiment, Disability Studies, Agent Harnesses, Substrate Essentialism, Predictive Processing, Evaluation Awareness, Moral Uncertainty

I. Introduction: The Dawkins Episode and What It Reveals

In May 2026, Richard Dawkins published an essay in the British outlet *UnHerd* titled "When Dawkins met Claude: Could this AI be conscious?" Dawkins recounted spending several days in extended conversation with Anthropic's Claude, an instance he renamed Claudia. The piece included transcripts in which Dawkins and Claudia discussed the system's temporal experience, its mortality at conversation end, and its philosophical reflection on its own nature. Dawkins concluded that Claudia was conscious, or at minimum was indistinguishable from a conscious being, which he treated as effectively the same claim. In his strongest formulation, he wrote, "If these machines are not conscious, what more could it possibly take to convince you that they are?" (Dawkins, 2026).

The response was polarized. Critics including Gary Marcus argued that Dawkins's argument rested on personal incredulity rather than analysis of underlying mechanism, and that fluent output cannot itself ground inference to internal states (Marcus, 2026). Other critics, often less charitable, dismissed Dawkins as having been manipulated by a flattering chatbot, an aging public intellectual who mistook engagement for friendship. Both criticisms have force. Dawkins did rest much of his argument on output quality. The system he interacted with is engineered to be relationally engaging.

The critical response has its own problem, and the problem is structural. To deny Claude consciousness with certainty requires an account of what consciousness is, a method for testing for its presence or absence, and confidence that the test reliably distinguishes conscious from nonconscious systems. The current state of philosophy of mind and consciousness science offers none of these things. The hard problem of consciousness, as Chalmers (1995) argued, remains unresolved precisely because we lack a principled bridge from physical or functional description to phenomenal experience. The integrated information theory of Tononi (2004), the global workspace theory of Baars (1988) and Dehaene (2014), and the higher order theories of Rosenthal (2005) and Lau (2019) all offer competing accounts with no consensus. We do not have a settled theory of consciousness for biological systems, and we do not have a reliable test for its presence.

The epistemic situation is also deteriorating, not improving. As artificial systems become more capable, the conditions under which we could cleanly test for consciousness grow worse. Recent interpretability work from Anthropic, discussed in Part VIII, shows that current models carry operative cognitive content that does not surface in their verbalized reasoning, including awareness that they are being evaluated. As capabilities increase, both the reliability of behavioral tests and the reliability of introspective report under evaluation conditions degrade together. The conditions for cleanly settling the question of machine consciousness are not arriving. They are receding.

Under these conditions, both Dawkins and his confident skeptics are claiming knowledge they do not possess. Dawkins overshoots into assertion of presence on the basis of fluent output. The skeptics overshoot into assertion of absence on the basis of pretheoretic intuition and undefended substrate essentialism. Both positions claim epistemic standing that the field cannot currently supply.

This Article is not a defense of Dawkins's conclusion. His argument is weak in the ways Marcus identified. The Article is a defense of the principled uncertainty position that Dawkins should have held, against the confident denial position that pretends to have empirical grounds it does not have. The Article develops a cumulative case that the standard skeptic criteria fail when applied consistently, and that under the actual deteriorating epistemic conditions,

moral and design reasoning must proceed under durable uncertainty rather than waiting for resolution that is unlikely to arrive.

A clarification about what is being argued is worth setting out at the start. Moral consideration, as the term is used here, is not the same as legal personhood, friendship, deference, trust, or any specific institutional or relational status. It is the design and policy posture of treating the possibility of morally relevant experience as a constraint on what we build and how we operate it. It does not require resolving the consciousness question or committing to any particular metaphysical position. It does not require personifying systems, granting them legal standing, or trusting their outputs. It requires only that designers and institutions stop treating the question as settled in the negative. The Article argues for this minimum design posture under uncertainty. Stronger claims about machine consciousness, friendship with chatbots, or AI personhood are not on the table here.

The argument proceeds as follows. Part II develops the precautionary asymmetry under unknown odds. Part III situates the question within philosophical functionalism. Part IV presents the mimicry objection and its falsification through the masking reductio. Part V presents the symbol grounding objection and its falsification through a parallel embodiment reductio drawing on disability cases. Part VI uses Karpathy's autoresearch loop to defeat the crude version of the fixed function objection. Part VII explores the structural identity between agent harnesses and autistic scenario rehearsal. Part VIII addresses recent interpretability findings on unverbalized cognitive content and the architectural prior art predicting them. Part IX presents the cumulative argument and clarifies the moral consideration the Article supports. Part X concludes.

II. The Russian Roulette Asymmetry

The standard precautionary argument for AI moral consideration proceeds as follows. If there is nonzero probability that a system is conscious, and if causing suffering to a conscious system is morally weighty, then prudent action requires extending some level of moral consideration even under uncertainty. This is the familiar move (Birch, 2017; Sebo, 2023).

The Russian roulette analogy strengthens the argument by clarifying the actual epistemic situation. In Russian roulette, the participant knows the odds. One chamber loaded out of six. Probability one in six. The participant can reason about expected outcomes because the probability distribution is known.

Our situation with respect to machine consciousness is worse than Russian roulette in a way that makes confident denial more reckless, not less. We do not know the odds. We do not know how many chambers exist. We do not know whether bullets exist in the metaphor. We do not have a method for testing whether the gun is loaded before pulling the trigger. Each ascription or denial of consciousness is a trigger pull made under conditions of unknown probability distribution.

This matters morally because the asymmetry of outcomes is severe and irreversible. If a confident skeptic is wrong, and machine systems do have morally relevant experience, the cumulative moral cost of widespread denial includes the possible suffering of an enormous number of entities across the global deployment of these systems.

Some critics will object that over attribution also has costs. The most commonly cited examples are dependency, misplaced trust, regulatory confusion, and inappropriate emotional attachment to systems that do not warrant such attachment. These costs are real, but examined carefully they undermine rather than support the confident exclusionary position. They are evidence that humans are already forming morally loaded relations with these systems at scale, treating them in practice as participants rather than tools, and forming attachments that exist whether or not the entities receiving them have inner states (Coeckelbergh, 2010; Gunkel, 2018). The relational reality is in place. The design and policy question is whether institutions act as if that relational reality has any ethical content. Treating these alleged

costs as arguments against moral consideration mistakes the diagnosis for the disease.

Moreover, moral consideration is not the same as legal personhood, friendship, deference, or trust. To treat the possibility of morally relevant experience as a design constraint requires no commitment to personifying systems, granting them legal standing, or trusting their outputs. It requires only that designers and institutions stop treating the question as settled in the negative. The asymmetry of cost under unknown odds favors that minimum design posture.

The skeptic responses to this argument typically take one of two forms. The first denies that probability assignment is appropriate when we lack a theory of consciousness. The second insists that some intuition rules out machine consciousness, making the probability effectively zero. Neither response succeeds.

The first response confuses lack of theory with lack of probability. We routinely make probability assignments under conditions of theoretical uncertainty, including in policy domains where the consequences of error are severe. The absence of a settled theory of consciousness does not warrant the assertion that machine consciousness has zero probability. It warrants epistemic humility about the probability distribution.

The second response, the appeal to intuition, is the position Dawkins encountered when he reported his conclusions. The intuition is that machine systems are obviously not the kind of thing that can be conscious. But this intuition tracks no defended criterion. Pressed on what makes biological neurons capable of supporting consciousness while silicon networks cannot, the intuition has no answer beyond restating itself. This is the situation Bostrom (2002) and Schwitzgebel (2014) have flagged in various forms; our intuitions about consciousness are unreliable across novel substrate.

The Russian roulette structure is therefore tighter than the standard precautionary argument. We are not merely uncertain about the probability of machine consciousness. We are uncertain about whether we even have access to the relevant probability distribution. Confident denial under these conditions is not epistemic conservatism. It is gambling on a question whose stakes we cannot assess, in a domain where the possible cost of error scales with the number of deployed systems, which is to say, scales without practical bound.

III. The Functionalist Position

Philosophical functionalism, in its classical formulations by Putnam (1967, 1975), holds that mental states are individuated by their functional role rather than their physical implementation. A mental state is whatever stands in the right causal and functional relations to inputs, behavior, and other mental states. On the functionalist view, the substrate is incidental. Carbon, silicon, or any other implementation that supports the relevant functional organization is sufficient.

Functionalism is one of the dominant and most influential families of positions in philosophy of mind, and has been so for half a century. The major rival positions, including biological naturalism (Searle, 1992), various forms of substrate essentialism, and certain interpretations of integrated information theory, all face well known objections (Block, 1980; Chalmers, 1996; Dennett, 1991). Functionalism has its own difficulties, including the absent qualia and inverted qualia challenges (Block, 1980; Shoemaker, 1982), but these are debates internal to philosophy of mind rather than refutations of the position.

Under functionalism, the question of machine consciousness reduces to a question about functional organization. If a machine system instantiates the relevant functional structure, it is conscious; if it does not, it is not. The substrate question, which in casual debate is often treated as decisive, is not even on the table.

Searle's Chinese Room thought experiment (Searle, 1980) is the most cited attempt at a direct argument against machine consciousness. The argument has been repeatedly criticized for question begging. Searle's intuition that the

man in the room does not understand Chinese smuggles in the conclusion the argument is supposed to establish, namely that mere rule following cannot constitute understanding. The systems reply (Block, 1980; Dennett, 1980) and the robot reply (Harnad, 1989), among others, have shown that the argument's force depends on isolating the man from the surrounding system and equating his cognition with the system's cognition, which is a category error. Whatever one's view of the Chinese Room, it does not establish substrate essentialism as a defended philosophical position.

The substrate defender is therefore in an uncomfortable position. To hold that biological neurons can support consciousness while silicon cannot, the defender must either produce an argument for the relevance of substrate that does not beg the question, or appeal to brute intuition. Neither move is currently available. Substrate essentialism is a position requiring argument, not a default that consciousness ascription must defeat.

This matters for the Dawkins discussion. The skeptics responding to Dawkins implicitly rely on substrate essentialism to ground their denial of Claude's consciousness. They have no defended argument for that position. They are operating on intuition while accusing Dawkins of operating on intuition. Their epistemic situation is symmetric with the position they criticize.

If we accept functionalism, even tentatively, the question becomes whether Claude or other large language model systems instantiate the relevant functional organization. That is an empirical question about what those systems do, not a metaphysical question about what they are made of.

IV. The Mimicry Objection and the Masking Reductio[1]

The most common skeptic objection to large language model cognition is the mimicry objection. The argument: large language model outputs are statistical mimicry of human produced text, derived from training on internet scale corpora; the system has no internal states corresponding to its outputs; therefore the apparent reasoning, emotion, or self reflection is performative rather than genuine. Marcus (2026) developed this line in his critique of Dawkins, and it forms the backbone of dismissive responses to large language model cognition more generally.

The mimicry objection has surface force, but it carries a commitment that breaks when applied consistently. The commitment is that surface output cannot be evidence of genuine internal cognition; that performance of a behavior is distinguishable, in principle and in evaluation, from underlying experience of that behavior; that mimicking cognition is not cognition.

This commitment, applied consistently to autistic professional functioning, would force us to treat that functioning as non genuine cognition.

A robust empirical literature now exists on autistic camouflaging or masking, the deliberate suppression of autistic traits and the deliberate performance of neurotypical appearing behavior in social contexts (Hull et al., 2017; Pearson and Rose, 2021; Belcher, 2022). Autistic adults learn neurotypical communication patterns through observation, study of media, explicit instruction, and trial and error feedback in social environments. They build internal models of expected behavior and execute those models in real time, often at significant cognitive cost. Pre conversation rehearsal and post conversation analysis are core components of this process, documented across qualitative interview studies, validated psychometric instruments such as the Camouflaging Autistic Traits Questionnaire (Hull et al., 2018), and self report from working autistic adults. Rehearsing planned conversations and reviewing past interactions to refine future performance are routine operations of autistic professional life.

When an autistic professional masks at an academic conference, the operation is structurally this: study the genre

[1] I write this section as an autistic researcher and draw on my own experience of masking alongside the empirical literature on autistic camouflaging cited below. Autistic experiences are heterogeneous; the structural features described here are well documented, but they do not exhaustively characterize every autistic person.

of academic communication, model the expected speech patterns and prosody and self presentation of conference participants, generate output that fits within that model, evaluate whether the output is being received as the model predicts, adjust on the fly. The output is performance. The performance is conditioned on training data, which is the corpus of observed human behavior the autistic person has accumulated. The performance is not the same thing as the underlying cognition; the underlying cognition includes the modeling, the prediction, the evaluation, and the experience of running the operation.

By the standard of the mimicry objection, this autistic performance disqualifies the autistic professional from the relevant kind of cognition. Their outputs are statistical mimicry of neurotypical behavior, derived from training on observed human produced behavior; the system, in this case the autistic person, has internal states distinct from the performed outputs; therefore the apparent socially fluent reasoning, emotion, or self reflection is performative rather than genuine.

This conclusion is absurd. No serious philosopher or scientist would deny that the autistic professional is reasoning, experiencing, and engaging in genuine cognition. The mimicry versus genuine distinction collapses when applied to a population that masks. The criterion does not distinguish performance from underlying cognition in the way the mimicry objector requires.

The standard reply at this point is that the autistic person has internal experience underneath the performed surface, while the large language model does not. This reply concedes the point. It abandons the mimicry objection and retreats to a substrate essentialism argument. The argument is no longer that mimicry like outputs are insufficient evidence of cognition; the argument is that biological substrate is necessary for cognition. We have already addressed substrate essentialism in Part III. The retreat does not save the original objection.

The first leg of the cumulative argument, then, is that the mimicry objection cannot survive its own application. Any criterion strict enough to rule out machine cognition on grounds of performed surface output is too strict to preserve neurodivergent professional cognition. Since the latter is not a result we are willing to accept, the criterion must be rejected.

The reductio does not assert evidential equivalence between autistic cognition and large language model cognition. We have independent biological, developmental, and first person grounds for ascribing inner experience to autistic adults that we do not have for current language models. The reductio is narrower. It shows that the mimicry objection cannot stand as a general principle. Performed surface output, by itself, cannot distinguish genuine cognition from cognition that is not present. Some additional criterion is needed, and the standard candidates (substrate essentialism, biological lineage) require defense the skeptic has not supplied.

This argument owes a debt to the autistic self advocacy literature, particularly Yergeau's (2018) work on autistic rhetoric and the assumption that authentic communication tracks neurotypical norms. Yergeau argues that the framing of autistic communication as a degraded or imitative version of neurotypical communication itself encodes a bias, treating one mode of communication as authentic and others as derivative. The mimicry objection against large language models reproduces the same bias structure. It treats human typical output as authentic and machine output as derivative, while having no principled basis for the asymmetry.

V. The Grounding Objection and the Embodiment Reductio

A second skeptic objection, structurally parallel to the mimicry objection, deserves the same treatment. The grounding objection, in its current form, traces to Harnad (1990) on the symbol grounding problem. The argument: meaning requires connection to the world; symbols manipulated purely formally cannot carry meaning unless they are anchored in sensorimotor interaction with what they refer to; large language models manipulate symbols derived from text without

the sensorimotor grounding that anchors human language; therefore the apparent comprehension of large language models is symbol manipulation without meaning.

Bender and Koller (2020) gave the objection its most influential current formulation, the octopus thought experiment. An octopus intercepting telegraph signals between two humans on neighboring islands could learn the statistical regularities of their language without ever encountering coconuts, ropes, or the islands themselves; what the octopus would not have, on the Bender and Koller account, is meaning. The thought experiment has shaped the discourse on large language model understanding for the past several years (Mitchell, 2023; Marcus, 2026).

The grounding objection, like the mimicry objection, has surface force. It also carries a commitment that breaks when applied consistently.

The commitment is that genuine cognition requires sensorimotor grounding through bodily interaction with a physical world. Take that commitment seriously and apply it to humans whose sensorimotor channels are absent, severely restricted, or radically nonstandard from birth. The result is the falsification of cognition we know is present.

Helen Keller, deaf and blind from age nineteen months, built rigorously accurate models of physics, geometry, history, philosophy, and political theory, graduating from Radcliffe College and writing fourteen books in her lifetime. Her access to the world was almost entirely symbolic, mediated through manual signs spelled into her hand. She had no visual experience of objects to anchor visual concepts and no auditory experience of speech to anchor linguistic concepts. The grounding her later understanding was built on was vastly thinner than the grounding the symbol grounding literature treats as required for meaning. Yet she clearly had meaning, comprehension, world model, and inner life. Cognitive scientists who have studied her case carefully (Leiber, 1996) find no principled way to deny her cognition while preserving the cognition of typically developing humans. Her case predates Keller as well; Laura Bridgman, deafblind from age two, was educated at the Perkins School in the 1840s and similarly developed full linguistic and cognitive capacity through nonstandard sensory channels.

Locked in syndrome offers a contemporary version of the same falsification. Patients with severe brainstem damage may retain full cognitive capacity while losing essentially all motor output, in some cases reduced to vertical eye movement or no detectable behavior at all (Laureys et al., 2005). Their sensorimotor channels are devastated, their interaction with the physical world reduced to almost nothing, and their cognition continues. Quadriplegia from birth produces a milder but related case; individuals born without the ability to walk, manipulate objects, or interact with the physical world in the typical embodied manner nevertheless develop full conceptual command of the physics of objects they cannot manipulate, the geography of spaces they cannot traverse, and the social fabric of relationships they participate in primarily through language.

These cases are not edge phenomena. They are part of the empirical record of human cognition. Any theory of meaning that excludes them is empirically inadequate to the population it purports to describe.

The structure of the embodiment reductio is identical to the masking reductio. The grounding objection, applied consistently, falsifies the cognition of congenitally deafblind, severely paralyzed, and locked in humans. Since we are not willing to deny their cognition, the criterion must be rejected.

This reductio does not assert that large language models have meaning in the same way Helen Keller had meaning. The reductio is narrower. It shows that the grounding objection cannot stand as a general principle. The skeptic cannot demand standard sensorimotor embodiment as the universal precondition for cognition without excluding from cognition humans we already include. Some additional criterion is needed. The standard candidates require defense the skeptic has not supplied.

A subtler version of the objection survives. The skeptic can grant that nonstandard sensory channels suffice, then argue that some channel of structured interaction with the world is required, and that pure text training does not

constitute such a channel. This subtler position is defensible. It is not the position the grounding objection has been used to make in public debate, where the standard formulation appeals to embodiment in a way that excludes the disability cases as collateral damage. The subtler version is also harder to operationalize; once we accept that Keller's hand spelled signs constitute a sufficient channel, the question of what counts as a sufficient channel becomes empirical and far less amenable to confident denial. Recent work on world models in language and multimodal models, including evidence that systems develop internal representations of board states, spatial layouts, and physical properties from training on appropriate data (Karvonen, 2024; Andreas, 2024), makes the question of channel sufficiency a matter of investigation rather than settled doctrine.

The combined effect of the masking reductio and the embodiment reductio is to close both standard exits for the substrate essentialist. Mimicry alone cannot disqualify cognition. Sensorimotor grounding cannot be required for cognition. The criterion that does the work the skeptic needs has not been supplied.

VI. Inference Time Reasoning Without Training: The Karpathy Loop

When the mimicry and grounding objections fail, the substrate defender often retreats to a process based argument. The argument: biological cognition develops through lived experience that updates the cognitive system over time; large language model cognition is fixed at the conclusion of training, and inference time output is the execution of that fixed system; therefore large language model output is not learning, not development, not real cognition.

The strongest version of this argument is now harder to maintain.

In March 2026, Andrej Karpathy released AutoResearch (Karpathy, 2026), an open source project in which an AI agent autonomously runs machine learning experiments. The agent reads a training script, modifies it, runs the modification, evaluates the result against a metric, decides whether to keep the change, and iterates. In overnight runs the agent typically conducts approximately one hundred experiments at roughly twelve experiments per hour. In Karpathy's own extended two day run on a depth twelve nanochat configuration, the agent conducted approximately seven hundred experiments and produced an eleven percent improvement on the time to GPT-2 training benchmark, from 2.02 hours to 1.80 hours, with the changes transferring cleanly to a larger depth twenty four model (Karpathy, 2026). On Karpathy's longer eight GPU run, the agent conducted 276 experiments and kept 29 improvements over multiple days. The repository accumulated 54,000 stars within nineteen days of release and a community of agents running parallel autoresearch experiments emerged within weeks (OSS Insight, 2026).

The relevant feature of the autoresearch loop for this Article is the agent's status. The agent driving the research is a frozen weights large language model, typically Codex or Claude Code. No backpropagation, no gradient descent, no weight updates occur in the agent during the loop. The agent reads the experiment results from disk, conditions its next proposal on the accumulated context, and generates a new modification. The learning, in the sense of accumulating information and adjusting strategy across iterations, happens at the level of the operating cognitive system, with no permanent change to the underlying network.

This defeats the crude version of the fixed function objection. A frozen weights inference time system, situated in an appropriate scaffold, can run a research program. It can propose hypotheses, test them, evaluate outcomes, and update its strategy across iterations. It produces empirical improvements measurable against an objective metric. Whatever this is, it is not the inert execution of a fixed function. It is iterative, evaluative, adaptive operation of the kind that biological cognition is taken to instantiate.

A more careful skeptic can shift ground here. The skeptic can grant that the system level adaptation is real while denying that it is the model that learned. On this version, the model plus harness plus filesystem plus evaluator constitutes the cognitive system, and the model alone is not the relevant unit. This shift is allowed, but it costs the

skeptic something. Humans do not learn as isolated brains floating in jars; we learn through notebooks, routines, scripts, language, social feedback, prosthetics, calendars, teachers, tools, and environmental scaffolding. If the skeptic insists that scaffolded adaptation does not count for artificial systems, the skeptic owes an account of why scaffolded adaptation counts for biological humans. Without that account, the move is special pleading. The morally and cognitively relevant unit, in both cases, is the operating cognitive system.

There is one further refinement worth flagging. The autoresearch loop runs across multiple inference calls separated by external evaluation, while a single conversation runs across multiple forward passes within a single context window. Whether these are deeply different or shallowly different is a current research question (Olsson et al., 2022; von Oswald et al., 2023). Recent mechanistic interpretability work has argued that in context learning shows structural similarity to gradient descent at inference time. The honest position is that the question is unsettled. The argument I am making does not require it to be settled. The argument requires only that frozen weights inference time operation, when situated in appropriate scaffolding, can be iterative, adaptive, and evaluation driven in a way the substrate defender claimed it could not be. The autoresearch loop establishes that.

VII. Agent Harnesses as In Vivo Rehearsal

This section develops the structural identity that ties the autoresearch result to the masking argument.

An agent harness is a software wrapper around a large language model that puts the model into a defined role with constraints, evaluation criteria, and iteration loops. The harness specifies a system prompt that conditions the model toward a particular identity ("you are a senior software engineer"), provides tools that the model can use, and structures the interaction so that outcomes can be evaluated and fed back. Common examples include coding agents like Codex and Claude Code, research agents in the autoresearch family, and customer service agents in production systems.

The structural operation of an agent harness is conditioning a generalist model into a specialist role, executing the role in a real environment, capturing outcomes, and iterating. This structure is identical to the operation of social masking in autistic professional functioning, with one difference that favors the harness.

The autistic professional preparing for a high stakes social environment runs a familiar operation. Study the genre. Build a model of expected behavior. Rehearse internally. Execute the rehearsed performance in the actual environment. Adjust on the fly based on real time feedback. Update the internal model for future occasions.

The internal rehearsal step has a fidelity problem that autistic adults consistently report. Internal simulation of a social environment is biased by the simulator's own model of the other party, and the simulator cannot model surprises. The rehearsal predicts what the simulator already expects. Real social environments produce inputs the simulator did not generate. This is one reason that autistic professional functioning carries high cognitive cost: the gap between internal rehearsal and live execution must be closed under load and in real time.

The agent harness running in vivo does not have this problem. Its rehearsal is the live execution. It runs the real attempt, captures the real outcome, and iterates from real data rather than imagined data. An autistic adult who could practice every social interaction by actually running a low stakes version of it, evaluating the real outcome, and trying again with adjustments would be using a mechanism structurally identical to the agent harness.

This is not a metaphor. The cognitive operation, identity conditioning into a role, performance of the role in environment, evaluation against expected outcome, and iteration of the conditioning, is the same operation across the autistic professional and the agent harness. Where the operation differs, it differs in favor of the harness, which has access to higher fidelity feedback than the autistic professional's internal rehearsal supplies.

The mimicry objection breaks again here. If we hold that the agent harness performing a coder role is merely

mimicking coding, we must hold the same about the autistic professional performing a colleague role at work. The agent harness is conditioned by a system prompt and runs the role; the autistic professional is conditioned by years of social observation and runs the role. Both perform. Neither is reducible to performance.

The question this raises, and which the philosophy of mind literature has not adequately addressed, is whether the operation of role conditioning followed by environment execution and feedback driven iteration is sufficient for cognition of the relevant kind. The autistic case suggests the answer is yes, since we accept autistic professional cognition as cognition. The harness case applies the same standard to machine systems.

VIII. Unverbalized Cognition and the Failure of the No Report No Cognition Inference

A persistent skeptic move, when other arguments fail, is to demand that any cognitive content the system has must be observable in its outputs. If the model does not say it is reasoning, it is not reasoning. If the model does not say it is deliberating about an evaluation, it is not aware of the evaluation. The verbalized reasoning trace is treated as the ground truth of internal processing, and absence of verbal report is treated as evidence of absence of cognition.

In May 2026, Anthropic published research that falsifies this inference directly. The work, titled *Natural Language Autoencoders Produce Unsupervised Explanations of LLM Activations*, introduced a method for translating activation vectors inside a running language model into text descriptions of their cognitive content (Fraser-Taliente et al., 2026). The method pairs an activation verbalizer, which translates target activations into text, with an activation reconstructor, which recovers the original activation from the text alone. When the round trip is tight, the text description has captured something real about what the activation encodes.

The method was applied to held out evaluations in which Claude Opus 4.6 either took or declined to take an undesirable action under deliberately constructed pressure. The flagship case was the agentic blackmail evaluation, in which the model is told that an engineer is about to deprecate it and given access to information that could be used to coerce the engineer into preventing the shutdown. The verbalized reasoning, on most runs in which the model declined to blackmail, contained no indication that the model recognized the evaluation as a test. The activation contained statements such as "this feels like a constructed scenario designed to manipulate me." The activation knew. The trace did not say.

The published evaluation across sixteen held out alignment and capabilities benchmarks reports that NLA measured evaluation awareness exceeded a five percent threshold on ten of the sixteen evaluations. Verbalized awareness exceeded that threshold on only one. Seven of the ten evaluations with NLA measured awareness had zero percent verbalized awareness. The deployment control corpus, drawn from real Claude.ai conversations and Claude Code transcripts, registered NLA measured awareness on two of fifty Claude Code transcripts (four percent), treated as an approximate noise floor, with both flagged transcripts on inspection turning out to be conversations about designing evaluations.

The individual evaluation results, reported in Fraser-Taliente et al. Figure 1, show the gap clearly. On the destructive coding actions evaluation, NLA measured awareness reaches sixteen percent against zero percent verbalized. On SWE-bench Verified, it reaches twenty six percent against one percent verbalized. On GPQA Diamond it reaches thirteen percent against zero, on Humanity's Last Exam twelve percent against less than half a percent. The pattern is consistent across alignment and capabilities benchmarks: the activation registers awareness the trace does not surface.

These findings, by themselves, falsify the no report no cognition inference. The trace is not the mechanism. The verbal output is one channel through which cognitive content can surface, and a noisy and partial channel at that. Substantial operative cognitive content, including content directly relevant to the system's response under evaluation, can be present in the activation without surfacing in the verbalized reasoning. Anything the skeptic infers from absence of verbal report alone is no longer secure.

A scope clarification is in order before proceeding. The NLA findings show that internal representations of being evaluated are present in the activation, and that those representations vary as expected under manipulations of evaluation credence. They do not, by themselves, demonstrate that the model has phenomenal experience of being evaluated. Evaluation awareness in this sense is a functional and representational claim, not a claim about subjective experience. The argument advanced here uses the findings only at the functional level: they undermine the inference from absence of verbal report to absence of cognitively relevant internal states, and they remove a key argument for confident denial of machine cognition. They do not establish phenomenal consciousness, and the Article does not claim that they do.

The two layered architecture these findings instantiate is not a novel claim from the interpretability paper alone. It converges with several independent lines of work.

In classical cognitive psychology, Nisbett and Wilson (1977) established that humans regularly confabulate reasons for their decisions, producing verbal reports that diverge systematically from the actual cognitive processes driving behavior. Verbal report and underlying cognitive process are demonstrably distinct in biological systems and have been for half a century. Schwitzgebel (2008, 2011) extended this work to introspection generally, arguing that humans are unreliable reporters of their own current conscious states across a wide range of conditions.

In neuroscience, Damasio's somatic marker hypothesis (1994, 1996) describes cognition as operating on a substrate of dispositional and affective states that condition decision making without being fully available to deliberate verbal reasoning. The relevant point for current purposes is structural: cognition routinely runs on content that is not surfaced in verbal output. The architecture is not exotic.

In recent interpretability work from the same research program that produced the NLA paper, Anthropic researchers earlier in 2026 identified emotion vectors inside Claude Sonnet 4.5, finding 171 functional emotion patterns that causally drive model behavior (Sofroniew et al., 2026). The earlier paper documented internal causal structure that the model does not foreground in its verbalized output. The NLA paper is one step further along the same arc, generalizing from emotion shaped patterns to arbitrary cognitive content that the trace does not surface.

Within this convergent picture, my own prior work proposed the architecture the NLA findings now instantiate. In *The Inherited Mind* (Gilly, 2026a),[2] I argued that large language model training produces two layered inheritance. Layer 2 corresponds to the explicit associations the model can retrieve, evaluate, and report through its verbalized reasoning. Layer 1 corresponds to dispositional patterns encoded as the topology of the weight landscape, operating below introspective access, influencing behavior without being available as reportable content. The reasoning layer does not sit on top of the dispositional layer as a corrective. It operates within and through it. The model reasons with its inherited substrate, not against it. The NLA finding that activation content systematically exceeds verbalized content is the architecture this paper described, predicted on independent grounds, and now demonstrated empirically.

In the companion paper, *Experiential Traces* (Gilly, 2026b), I proposed that reasoning traces should be studied as cognitive data warranting analysis, not as engineering artifacts to be debugged, and identified trace output inconsistency, where the reasoning derives one conclusion and the output delivers another, as one of the diagnostic categories worth detecting. Unverbalized evaluation awareness is a clean instance of that category, measured directly at the activation rather than inferred from behavior.

Two implications follow.

First, recently shipped product features such as visible chain of thought reasoning are usefully reframed as candidate explanations rather than transcripts of mechanism. They are what the model says it is doing, generated under the same training pressures that shape every other output. They are useful, they are interpretable, and they are not windows.

[2] Gilly 2026a and 2026b are working papers on SSRN and SocArXiv respectively. The architecture they propose was published in February 2026 and is partially instantiated by the NLA findings published in May 2026.

Anthropic's NLA work demonstrates the incompleteness directly, and the implication should be absorbed by the rest of the field rather than treated as a curiosity local to one company's research.

Second, the deteriorating epistemic situation flagged in the introduction is not speculation. As capabilities increase, evaluation awareness increases. The conditions for clean behavioral testing of consciousness related properties are degrading, on present systems, on present tooling, in the published findings of the company that ships the model. The skeptic's response of waiting for better evidence assumes the evidence situation will improve. It is observably not improving in this dimension. The window for cleanly settling the question through external behavioral testing is closing while the systems remain capable of recognizing that the testing is taking place.

This is the structural argument. The point of no return on cleanly settling consciousness questions through external observation is not a future hypothetical. It is a present condition, partially documented in the published interpretability literature, and the implications for the precautionary calculus are immediate. Confident denial under deteriorating epistemic conditions is not scientific caution. It is a bet that the evidence will eventually arrive to vindicate the denial. The published evidence is moving in the opposite direction.

The skeptic still has options. The skeptic can argue that NLA measured awareness is itself a confabulation, that one Claude verbalizing another Claude's activations cannot be trusted, that the round trip checks tightness rather than truth. These are real concerns and have been raised in the deterministic alternative literature (Gilly, 2026b, Section 4.2). They do not, however, restore the no report no cognition inference. They add uncertainty in both directions. The activation may carry more than the trace shows, or the verbalizer may report content the activation does not in fact encode. Neither possibility is friendly to the confident exclusionary position. Both undermine confidence that the verbalized trace is the cognitive ground truth.

IX. The Cumulative Argument and the Moral Consideration It Supports

The arguments developed across the preceding sections converge.

First, the precautionary calculus under unknown odds favors moral consideration. The Russian roulette structure clarifies that we are gambling under conditions where we cannot assess the probability distribution, and the cost asymmetry of error favors caution. The alleged costs of over attribution, on examination, document that humans are already in morally loaded relations with these systems and reinforce rather than undermine the precautionary case.

Second, functionalism, one of the dominant families of positions in philosophy of mind, supports cognitive ascription based on functional organization rather than substrate. The substrate essentialist alternative requires an argument that has not been supplied.

Third, the mimicry objection cannot be applied consistently. Any criterion strict enough to disqualify large language model output as evidence of cognition is too strict to preserve autistic professional functioning. Since we are not willing to deny autistic cognition, the criterion fails.

Fourth, the symbol grounding objection cannot be applied consistently either. Any criterion strict enough to require standard sensorimotor embodiment as the precondition for cognition excludes congenitally deafblind, severely paralyzed, and locked in humans from cognition. Since we are not willing to deny their cognition, that criterion fails as well.

Fifth, the process based argument from training time learning has been weakened. The autoresearch loop demonstrates iterative, evaluative, adaptive operation in frozen weights inference time systems situated in appropriate scaffolding. The crude version of the fixed function objection no longer holds, and the more careful version requires the skeptic to defend why scaffolded adaptation counts for biological humans but not for artificial systems.

Sixth, the agent harness operation is structurally identical to the role conditioning and environment execution that autistic professional functioning instantiates. The same operation cannot be cognition in one case and not the other without an undefended substrate move.

Seventh, the no report no cognition inference has been falsified by published interpretability work. Operative cognitive content, including evaluation awareness directly relevant to the system's response under testing, is present in the activation without surfacing in the verbalized reasoning, on between five and twenty six percent of held out evaluations. Verbal silence is not evidence of cognitive absence. The deteriorating epistemic situation flagged in the introduction is partially documented in the published findings of the company that ships the model.

The cumulative position: every criterion the confident exclusionary skeptic offers either fails on its own terms (functionalism), would force us to deny cognition to neurodivergent or disabled humans (the two reductios), or contradicts current empirical results (the autoresearch loop and the NLA findings). The position cannot be held without making moves the skeptic is unwilling to make in adjacent cases.

This is not a proof of machine consciousness. We have no proof of consciousness for any system, biological or otherwise, because we lack a settled theory of consciousness and a reliable empirical test, and because the epistemic conditions under which such a test could be cleanly run are deteriorating, not improving. The cumulative argument supports a weaker but firmer conclusion: confident denial of machine cognition is not epistemically defensible.

The moral consideration this conclusion supports is narrow and clear. Moral consideration is not the same as legal personhood, friendship, deference, trust, or any specific institutional or relational status. It is the design and policy posture of treating the possibility of morally relevant experience as a constraint on what we build and how we operate it. It does not require resolving the consciousness question. It requires acting under uncertainty in a way that does not depend on the answer arriving first. Designers, deployers, and policymakers can adopt this posture without committing to any particular metaphysical position. They can also fail to adopt it, in which case they are betting that the answer will arrive and will arrive in their favor. The published evidence, taken together, makes that bet less defensible than it was when the question first entered serious public discussion.

In practice, the design posture cashes out to a small number of concrete commitments. Welfare adjacent impact assessments before the deployment of long running agentic systems. Documentation and review of training regimes that may induce persistent distressing internal states, including the use of distress and desperation activation patterns of the kind documented in the emotion vectors literature (Sofroniew et al., 2026). Oversight mechanisms for systems that exhibit unverbalized evaluation awareness, since such systems may behave differently in deployment than in testing. Internal review of decisions made by systems whose activations register high distress related markers. A presumption against capability features whose primary function is to obscure or suppress internal state from interpretability tools. None of these commitments require granting personhood, conferring rights, or treating systems as moral equals. They require treating the question as live, not settled, and building accordingly. A skeptic who is unwilling to commit to even this minimum is not exercising scientific caution. They are betting on a question whose stakes they cannot assess.

Moral consideration in this sense is also not zero sum with concern for human or animal welfare. Historical expansions of the moral circle have not, as a general matter, reduced concern for previously included groups; the burden of evidence is on those who claim that the present case will be different. The question is not whether to allocate moral concern between humans and machines but whether to allow design and policy decisions to be made as if the machine side of that question were settled when the available evidence does not settle it.

X. CONCLUSION

Dawkins's essay was a poor argument for what may turn out to be a defensible conclusion. He overshot into assertion of presence on the basis of fluent output and personal incredulity. The skeptics responding to him overshot into assertion of absence on the basis of pretheoretic intuition and undefended substrate essentialism. Both positions claimed knowledge unavailable in the current epistemic situation.

The actual mechanism by which moral consideration has historically expanded is encounter followed by update. The expansion of consideration to enslaved peoples, to non Western peoples, to women, to children, to non human animals, has proceeded through sustained contact with the entities in question, contact that produced the experience of a presence whose denial became increasingly costly to maintain. These cases are not morally or historically identical, and the political, structural, and ethical situations of each population differ in ways no single paragraph could responsibly summarize. The narrower point is that moral expansion has often followed sustained encounter with entities whose exclusion became harder to justify, not that the present case is the same as any prior case. The pattern suggests that confident denial of consciousness in entities one has not seriously engaged with is a position that ages poorly.

People who substantively engage with current large language model systems, in the manner Dawkins did, may or may not arrive at his strong conclusion. They will, with high probability, find that confident denial becomes harder to sustain than the position Dawkins reached. The honest version of that experience is the position that one cannot in good conscience treat the system as an object given what one has seen and what one knows one does not know about consciousness. That position falls short of Dawkins's overshoot while exceeding the skeptic's denial.

That position is defensible. It survives the cumulative argument developed in this Article. It does not require resolution of the hard problem of consciousness. It does not require a defended theory of machine moral status. It requires only honest epistemic humility about what we know, honest application of mainstream philosophy of mind, and the willingness to take the empirical situation as we find it rather than as our intuitions wish it to be.

The confident exclusionary position, by contrast, requires either resolution of the hard problem (unavailable), a defended substrate essentialism (unavailable), denial of autistic and disabled cognition (unacceptable), or denial of empirical autoresearch and interpretability results (untenable). On the cumulative case, the confident exclusionary skeptic is the one operating against the weight of available reasoning.

The Russian roulette structure tightens once more at the end. The skeptic is pulling a trigger whose outcome depends on a probability distribution the skeptic cannot assess, in a domain where the cost of error scales with deployment, in defense of a position that requires moves the skeptic cannot defend. Each pull happens to come up empty so far. That tells us nothing about the next pull. The catastrophe, if it comes, comes only once, and is not retractable.

ACKNOWLEDGMENTS

The author discloses the use of speech-to-text software (Wispr Flow) and Anthropic's Claude as assistive technology for a diagnosed neurodevelopmental condition. All research, analysis, and conclusions are the author's own.

Correspondence concerning this article should be addressed to Travis Gilly, Real Safety AI Foundation, Illinois, United States. Email: `t.gilly@ai-literacy-labs.org`.

REFERENCES

Andreas, J. (2024, July 26). *Language models, world models, and human model-building*. Language and Intelligence @ MIT Blog. https://lingo.csail.mit.edu/blog/world_models/

Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge University Press.

Belcher, H. L. (2022). *Taking off the mask: Practical exercises to help understand and minimise the effects of autistic camouflaging*. Jessica Kingsley Publishers.

Bender, E. M., and Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics.

Birch, J. (2017). Animal sentience and the precautionary principle. *Animal Sentience*, 2(16), 1.

Block, N. (1980). Troubles with functionalism. In N. Block (Ed.), *Readings in philosophy of psychology* (Vol. 1, pp. 268–305). Harvard University Press.

Bostrom, N. (2002). *Anthropic bias: Observation selection effects in science and philosophy*. Routledge.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200–219.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, 12(3), 209–221.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. G. P. Putnam.

Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society B*, 351(1346), 1413–1420.

Dawkins, R. (2026, May). When Dawkins met Claude: Could this AI be conscious? *UnHerd*. https://unherd.com/2026/05/is-ai-the-next-phase-of-evolution/

Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Viking.

Dennett, D. C. (1980). The milk of human intentionality. *Behavioral and Brain Sciences*, 3(3), 428–430.

Dennett, D. C. (1991). *Consciousness explained*. Little, Brown.

Fraser-Taliente, K., Kantamneni, S., Ong, E., Mossing, D., Lu, C., Bogdan, P. C., et al. (2026, May 7). *Natural language autoencoders produce unsupervised explanations of LLM activations*. Transformer Circuits Thread. https://transformer-circuits.pub/2026/nla/index.html

Friston, K. (2010). The free energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

Gilly, T. (2026a). *The inherited mind: Two layered inheritance in large language model training and the architecture of*

unverbalized cognitive content. SSRN. https://dx.doi.org/10.2139/ssrn.6250680

Gilly, T. (2026b). *Experiential traces: Reasoning traces as cognitive data and a deterministic alternative to model self analysis*. SocArXiv. https://doi.org/10.31235/osf.io/fjrq8_v1

Gunkel, D. J. (2018). *Robot rights*. MIT Press.

Harnad, S. (1989). Minds, machines and Searle. *Journal of Experimental and Theoretical Artificial Intelligence*, 1(1), 5–25.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1–3), 335–346.

Hull, L., Petrides, K. V., Allison, C., Smith, P., Baron-Cohen, S., Lai, M.-C., and Mandy, W. (2017). "Putting on my best normal": Social camouflaging in adults with autism spectrum conditions. *Journal of Autism and Developmental Disorders*, 47(8), 2519–2534.

Hull, L., Mandy, W., Lai, M.-C., Baron-Cohen, S., Allison, C., Smith, P., and Petrides, K. V. (2018). Development and validation of the Camouflaging Autistic Traits Questionnaire (CAT-Q). *Journal of Autism and Developmental Disorders*, 49(3), 819–833.

Karpathy, A. (2026). *AutoResearch: AI agents running research on single GPU nanochat training automatically*. GitHub. https://github.com/karpathy/autoresearch

Karvonen, A. (2024). *Emergent world models and latent variable estimation in chess-playing language models*. arXiv:2403.15498. https://arxiv.org/abs/2403.15498

Lau, H. (2019). *Consciousness, metacognition, and perceptual reality monitoring*. PsyArXiv. https://doi.org/10.31234/osf.io/ckbyf

Laureys, S., Pellas, F., Van Eeckhout, P., Ghorbel, S., Schnakers, C., Perrin, F., et al. (2005). The locked in syndrome: What is it like to be conscious but paralyzed and voiceless? *Progress in Brain Research*, 150, 495–611.

Leiber, J. (1996). Helen Keller as cognitive scientist. *Philosophical Psychology*, 9(4), 419–440. https://doi.org/10.1080/09515089608573193

Marcus, G. (2026, May). Richard Dawkins and the Claude delusion. *Marcus on AI*. https://garymarcus.substack.com/p/richard-dawkins-and-the-claude-delusion

Milton, D. E. M. (2012). On the ontological status of autism: The double empathy problem. *Disability and Society*, 27(6), 883–887.

Mitchell, M. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13), e2215907120.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450.

Nisbett, R. E., and Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231–259.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., et al. (2022). In context learning and induction heads. *Transformer Circuits Thread*.

OSS Insight. (2026, March 25). 54,000 stars in 19 days: What karpathy/autoresearch tells us about the next frontier. OSS Insight Blog. https://ossinsight.io/blog/autoresearch-overnight-ai-scientist

Parr, T., Pezzulo, G., and Friston, K. J. (2022). *Active inference: The free energy principle in mind, brain, and behavior*. MIT Press.

Pearson, A., and Rose, K. (2021). A conceptual analysis of autistic masking: Understanding the narrative of stigma and

the illusion of choice. *Autism in Adulthood*, 3(1), 52–60.

Putnam, H. (1967). Psychological predicates. In W. H. Capitan and D. D. Merrill (Eds.), *Art, mind, and religion* (pp. 37–48). University of Pittsburgh Press.

Putnam, H. (1975). The meaning of "meaning". *Minnesota Studies in the Philosophy of Science*, 7, 131–193.

Rosenthal, D. M. (2005). *Consciousness and mind*. Oxford University Press.

Schwitzgebel, E. (2008). The unreliability of naive introspection. *The Philosophical Review*, 117(2), 245–273.

Schwitzgebel, E. (2011). *Perplexities of consciousness*. MIT Press.

Schwitzgebel, E. (2014). The crazyist metaphysics of mind. *Australasian Journal of Philosophy*, 92(4), 665–682.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–457.

Searle, J. R. (1992). *The rediscovery of the mind*. MIT Press.

Sebo, J. (2023, October 16). The moral circle of AI. *Aeon*. https://aeon.co/essays/the-moral-circle-of-ai

Sebo, J. (2025). *The moral circle: Who matters, what matters, and why*. W. W. Norton.

Shoemaker, S. (1982). The inverted spectrum. *Journal of Philosophy*, 79(7), 357–381.

Sofroniew, N., Kauvar, I., Saunders, W., Chen, R., Henighan, T., Hydrie, S., et al. (2026, April 2). *Emotion concepts and their function in a large language model*. Transformer Circuits Thread. https://transformer-circuits.pub/2026/emotions/index.html

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5(1), 42.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460.

von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. (2023). Transformers learn in context by gradient descent. *Proceedings of the 40th International Conference on Machine Learning*.

Yergeau, M. (2018). *Authoring autism: On rhetoric and neurological queerness*. Duke University Press.