The Ontic-Epistemic Distinction: Implications for Robust General Intelligence

Shreya Ishita[1]

March 2026

Abstract

The current pursuit of robust machine intelligence is largely predicated on a substrate independent, computational functionalist view of cognition, where sufficiently complex computational processing is expected to eventually yield generalized reasoning. This paper explores the ontological distinctions between these computational frameworks and biological cognition, specifically how these differences impact the capacity for semantic understanding. By analyzing phenomena such as the "reversal curse" where models fail to generalize the symmetry in identity relations (A=B implies B=A), and performance on novel reasoning benchmarks (e.g., ARC-AGI), this paper examines whether current model limitations are transient artifacts of scale or indicative of a distinct architectural category. Integrating Stevan Harnad's "symbol grounding problem" with Evan Thompson's biological model of "intrinsic normativity," I investigate whether robust general intelligence might require sense-making: a process distinct from information processing, whereby an agent's internal states are causally coupled with its environment via survival or system-wide stakes which grounds symbols in meaning. Current Large Language Models (LLMs) appear to lack this intrinsic normativity, and consequently may operate primarily as epistemic instruments rather than ontic agents. By introducing the concept of "ontic grounding", this paper presents a potential framework for distinguishing between the simulation of reasoning and true understanding, which could have implications for AI safety and governance.

[1]*All analyses of AI architectures and failure modes are based on publicly available data and open-source benchmarks.*

1. Introduction

The current trajectory of artificial intelligence research is largely predicated on achieving artificial general intelligence (AGI) via the scaling hypothesis, the empirical observation that model performance scales as a power-law function of compute, dataset size, and parameter count (Kaplan et al., 2020). Implicit in this paradigm is a substrate-agnostic, computational functionalist commitment regarding the nature of general intelligence, defined here as the capacity to flexibly adapt to solving novel problems across diverse domains. Since human cognition as the paradigm case of general intelligence achieves robust generalization through semantic understanding, constructing causal models of the entities underlying the sensory data (Lake et al., 2017), proponents of AGI are implicitly positing that sufficient statistical complexity can serve as a substitute for this semantic grounding. They operate on the premise that the "hallucinations" and generalized reasoning failures of current Large Language Models (LLMs) are not architectural deficits, but transient artefacts of insufficient scale.

However, as models have scaled from billions to trillions of parameters, a persistent tension has emerged between performance metrics and reliability. While frontier LLMs have achieved near-human or super-human fluency in diverse domains, they continue to exhibit fundamental brittleness in areas requiring causal reasoning and truth-tracking. This divergence between the appearance of competence and the reality of robustness is perhaps best exemplified by the “reversal curse,” where autoregressive models fail to generalize the logical symmetry of identity relations (Berglund et al., 2024), and by their plateauing performance on novel reasoning benchmarks that cannot be consistently solved via memorization alone (Chollet, 2019). Despite the overlay of reinforcement learning from human feedback (RLHF) to align model outputs, these systems continue to operate primarily via probabilistic sequencing (Bender et al., 2021), stitching together linguistic forms with high probability but without access to the communicative intent or causal constraints that ground those forms.

This paper argues that this robustness tension is unlikely to be resolved simply by increasing computational scale. I posit that the persistence of these errors could indicate an explanatory gap in the current paradigm. Drawing on the distinction between epistemic information (syntactic, copyable data about the world) and ontic information (the intrinsic, causal state of a system relative to its environment), I present a hypothesis of “ontic grounding” as a plausible structural condition for robust general intelligence. It explores whether capacities such as causal reasoning and truth-tracking, require a specific system-level causal relationship between an agent and its environment, a relationship that current deep learning architectures appear to lack.

Potential implications from this distinction suggest that the stakes are not merely theoretical, but could be urgently practical. These risks could arise precisely because epistemic fluency often presents itself as understanding. If we fail to distinguish between the simulation of reasoning (epistemic) and the possession of understanding (ontic), we could face several possible risks, of which two categories are discussed in this paper. First, an engineering risk: deploying systems in high-stakes environments (e.g., healthcare, law) that possess the fluency to persuade but lack the grounding to verify truth, leading to "silent failures" or automation bias. Second, an ethical risk: falling into what Shevlin (2024) describes as "unironic mentalising," where society grants moral status to systems based on behavioral simulation, thereby decoupling moral worth from the ontic reality of the subject and potentially creating a competition for moral consideration between human beings and non-sentient artefacts.

To advance this argument, this paper proceeds as follows. In Section 2, I review the current landscape of AI evaluation, arguing that functionalist approaches to measuring competence risk succumbing to "gerrymandering" or passing tests without possessing the underlying capacity. In Section 3, I formally introduce the concept of "ontic grounding" and lay out a hypothesis, synthesising Harnad's (1990) symbol grounding problem with Thompson's (2010) intrinsic normativity to offer a potential criterion for semantics and general intelligence. Section 4 anticipates and defends against objections from the computationalist view, specifically regarding simulation and multimodality. Finally, Section 5 explores the ethical implications, arguing that recognizing the ontic grounding gap could allow us to address Shevlin's "inconsistent triad" regarding the moral status of AI, preserving a principled distinction between human subjects and algorithmic tools.

2. Syntactic Scaling vs. Semantic Robustness

2.1. The Scaling Hypothesis

To evaluate the limits of artificial intelligence, one must first account for an empirical reality of the field: the effectiveness of scaling. Proponents of the scaling hypothesis argue that the functional deficits of current models are not categories of impossibility, but merely functions of insufficient computation. Kaplan et al. (2020) demonstrated that test loss (prediction error) declines predictably with compute, while Wei et al. (2022) identified "emergent abilities", which is the capacity to perform tasks such as arithmetic, translation, and chain-of-thought reasoning, that appear discontinuously only after models pass certain parameter thresholds. The philosophical implication of this engineering observation is often interpreted as empirical

support for a strong form of computational functionalism, the view that mental states like reasoning are defined solely by their input-output functions that can be computed. In the context of LLMs, this implies that reasoning, and the semantic understanding it entails, is a substrate-independent process that can be fully realized via sufficiently complex syntactic manipulation.

2.2. The Persistence of Asymmetry in the "Reversal Curse"

However, recent investigations into the logical structure of LLM knowledge reveal a flaw in this narrative. If scaling were truly bridging the gap between syntax and semantics, we should expect models to internalize the fundamental logical properties of the entities they represent, such as the symmetry of identity (if A=B, then B=A). The "reversal curse" (Berglund et al., 2024) indicates that autoregressive models exhibit a persistent failure to generalize this symmetry. A model trained on "Tom Cruise's mother is Mary Lee Pfeiffer" does not automatically infer "Mary Lee Pfeiffer's son is Tom Cruise." Critics might object that this finding is outdated, noting that state-of-the-art models (like GPT-4o or Claude 3.5) often succeed at such reversals. However, this objection risks confusing data augmentation with logical derivation. Recent models succeed at reversals involving famous figures not necessarily because they have learned the logical rule of symmetry, but possibly because their training corpora has been expanded to include the fact in both directions (Allen-Zhu and Li, 2024).

When tested on novel logical relations such as fictional facts introduced for the first time during the testing phase, where the model cannot rely on memorization, the failure seems to persist regardless of model size, at least in current architectures. This strongly suggests that the system is not navigating a semantic territory (where relations are intrinsic and multi-directional), but is rather traversing a syntactic map (where relations are statistical and directional). Scaling appears insufficient in fully resolving the reversal curse, because it possibly merely hides it behind a denser layer of memorized tokens. The system mimics understanding, but structurally retains the asymmetry of a "one-way street," a flaw incompatible with genuine general intelligence.

2.3. Evidence from ARC-AGI

If the reversal curse indicates a failure in logical symmetry, the performance of LLMs on the Abstraction and Reasoning Corpus (ARC-AGI) demonstrates a parallel failure in adaptive generalization. Proposed by Chollet (2019), ARC rigorously distinguishes between priors-heavy

tasks (which can be solved by memorizing vast amounts of training data, e.g., coding or translation) and priors-poor tasks (novel logic puzzles requiring on-the-fly abstraction).

While recent frontier models have achieved super-human scores on memorization-heavy benchmarks like MMLU, their performance on ARC-AGI has notably plateaued, lagging behind human capability (Chollet, 2024). This divergence empirically undermines the functionalist assumption that reasoning is merely a downstream effect of scale. It suggests that improved performance on memorization heavy benchmarks does not necessarily reflect gains in semantic understanding. The model can recall the history of the French Revolution (memorization) but fails to deduce simple physical rotation rules for a grid it has never seen (logical derivation).

This reinforces the hypothesis that scaling laws apply primarily to the interpolation of the training distribution, but yield diminishing returns on extrapolation. Without the semantic grounding to understand why a rule exists (e.g., object permanence or gravity), the system may lack the capacity to flexibly apply that rule in novel contexts; it can only simulate the rule if it appears in its syntactic history.

2.4. Architectural Gerrymandering

Current approaches to evaluating models are vulnerable to what Henry Shevlin (2021) identifies as the "specificity problem" in assessing non-human consciousness: if we rely solely on behavioral markers, we risk validating "gerrymandered" systems, which are machines engineered to pass the test without possessing the underlying capacity.

Contemporary LLMs may be viewed as the ultimate realization of Shevlin's warning. By optimizing models to produce specific output tokens (e.g., forcing a chain of thought structure), developers can unintentionally produce the appearance of reasoning without the architectural reality of it. The persistence of hallucinations and brittleness suggests that these interventions primarily serve as epistemic adjustments, modifying the surface behavior to mimic reliability, while the system remains architecturally disconnected from the causal structures it simulates. To address this, I propose we need to look beyond the behavior of the system to the ontic conditions of its architecture.

3. The Ontic Grounding Hypothesis

If current architectures are merely producing the appearance of reasoning through statistical mimicry, a question that pertains to this is, what could be the missing architecture feature required for genuine competence. To investigate this, we must look beyond the domain of

computer science to the philosophy of mind and cognitive science. I propose that the barrier preventing LLMs from achieving robust general intelligence is likely not computational, but ontological; specifically, a lack of ontic grounding.

3.1. Syntax Without Semantics

The foundational diagnosis of the LLM condition was provided by Stevan Harnad (1990) in his formulation of the "symbol grounding problem." Harnad argued that a system manipulating symbols based solely on their shape (syntax) can never derive the meaning (semantics) of those symbols, no matter how complex the rule-book. He illustrated this with the metaphor of the "Chinese dictionary-go-round": learning Chinese as a second language using only a Chinese-to-Chinese dictionary. One can define a word using other words, but without a point of entry from the external world (e.g., a sensory image of a horse linked to the symbol "horse"), one remains trapped in an infinite regress of empty signifiers.

Contemporary LLMs may operate as the ultimate scaling of the dictionary-go-round. They are trained on an effectively unbounded corpus containing nearly all human symbolic output. However, as Bender and Koller (2020) argue in their "Octopus" thought experiment, increasing the size of the dictionary does not solve the grounding problem; it merely expands the available syntax. The system possesses a detailed map of the statistical distribution of tokens (epistemic information), but lacks any causal connection to the referents of those tokens (ontic reality).

3.2. From Information Processing to Sense-Making

A question Harnad's symbol grounding problem raises is, if symbols cannot ground symbols, what can. Evan Thompson (2010), drawing on the deep continuity thesis between life and mind, argues that genuine cognition is not information processing (the manipulation of abstract symbols), but sense-making (the regulation of an organism's coupling with its environment).

In Thompson's view, meaning is not a dictionary definition; it is a relationship of value relative to survival, resulting in intrinsic normativity. This normativity arises in all autopoietic (self-creating) systems: entities that must actively maintain their structural integrity against decay. For a bacterium, a sucrose gradient is not merely data; it is food. The signal possesses intrinsic meaning because it carries system-level consequences such that if the bacterium misinterprets the signal (e.g., mistaking poison for food), it could cease to exist.

This logic of sense-making extends directly to human cognition, which represents a higher-order elaboration of this basic biological value system. For a human engaging in symbolic reasoning

(such as language or mathematics), intrinsic normativity expands to include social and sensorimotor dimensions. As Harnad (1990) argues, the meaning of elementary symbols (e.g., "jump," "apple") is anchored in the organism's ontic interaction with the physical world, and the invariant features detected by a living body. Even abstract concepts (e.g., "mathematics," "justice") are constructed from these grounded primitives via conceptual mapping (Lakoff and Núñez, 2000).

Critically, this does not imply that only biological systems can instantiate normativity. However, it does require that the system possess a causal architecture where grounding imposes consequences. If a human misinterprets the word "fire," they may burn; if they misuse social language, cooperation collapses. The meaning is enforced by the reality of the agent's existence. In contrast, current AI is allopoietic (made by others), whereby its internal state is epistemic: a symbolic representation (voltage) that is arbitrary relative to the physical machine running it. A hallucination in an LLM carries no structural or social consequence for the system itself; it only results in a gradient update. Because the system has no ontic stake in the accuracy of its symbols, it optimizes for the probability of the token, not the reality of the referent.

3.3. The Ontic Grounding Hypothesis

Synthesizing Harnad's critique of symbolism with Thompson's intrinsic normativity, I introduce the ontic grounding hypothesis. Rather than viewing grounding as a binary biological trait, I present it as a plausible structural condition for robust intelligence:

> Ontic Grounding is plausibly instantiated when an information processing system's internal states are not merely self-referential tokens, but are constrained by a direct, causal coupling with the environment such that errors in representation result in system-level consequences for persistence or goal achievement.

Consequently, we can distinguish between two types of intelligence:

1. **Epistemic intelligence (simulation):** The manipulation of representations based on syntactic rules or statistical probabilities. This scales with compute (the scaling hypothesis) but is prone to "hallucination" because the map can be altered without checking the territory.
2. **Ontic intelligence (sense-making):** The navigation of the environment based on a direct experience of grounded invariant features. This is robust because it is truth-tracked via its

reliance on the reality of the referent, such that a system cannot hallucinate a physical action without immediate causal feedback.

The robustness tension identified in Section 2 could be plausibly explained by the ontic grounding gap. The reversal curse and ARC-AGI failures persist as failures of generality likely because current LLMs operate primarily within epistemic intelligence, without the ontic architecture to ground them in semantics. Understanding the invariant causal rules of a system (e.g., 'falling') allows for infinite generalization across novel contexts. Syntactic correlations, no matter how scaled, remain brittle because they rely on probability distributions of past data. Thus, the failure of robustness is plausibly attributable to this lack of semantic grounding.

4. Why Simulation May Not Instantiate Ontic Grounding

The argument presented thus far relies on a distinction between epistemic simulation, creating an appearance of competence, and ontic grounding. A proponent of the current paradigm might object that this distinction is theoretically suspect. Two primary objections inevitably arise: first, the functionalist objection that a sufficiently detailed simulation of a system produces the same mental capacities as the system itself; and second, the multimodal objection that adding sensory data such as vision or robotics resolves the grounding problem. I address each in turn.

4.1. The Simulation Fallacy

The strongest objection from computational functionalists is that functional organization, defined here as the interplay between inputs, internal organization, and outputs, is sufficient for the production of mental states. A computational model such as an advanced LLM that can predict the next token in a causal sequence with high fidelity, could effectively capture the causal structure of reality. As Chalmers (1996) and others have argued via the principle of organizational invariance, if we can replicate the causal topology of a cognitive system, we should replicate its mental properties. In this view, the critique of “stochastic parrots” is simply a critique of insufficient resolution; a perfect parrot would be indistinguishable from a speaker.

However, this objection falls victim to what Seth (2025) and Thompson (2010) identify as the confusion between simulation and instantiation. To borrow a classic analogy used by both authors: a perfect computer simulation of a rainstorm captures the dynamic relations between pressure, humidity, and temperature, yet it does not make the computer wet. The simulation contains the epistemic profile of the storm (information about it), but lacks the ontic reality of the storm (the water itself). While Functionalism may adequately describe "undemanding" cognitive states (Shevlin et al., 2025) like knowledge, or performing calculations (where the

simulation of 2+2 is identical to the calculation of 2+2), that can be sufficiently encapsulated by their input output function, it fails when applied to "demanding" cognitive states like true understanding or robustness, which plausibly require the structure of “ontic grounding”.

Critics might rejoin that logical systems (such as symbolic solvers or physics engines) are built purely on symbols, and yet possess “truth” because they can follow causal rules rigidly. However, this risks conflating syntactic consistency with semantic grounding. A physics engine that calculates F=ma forces the system to behave as if it observes gravity, but the system remains ontically disconnected; it would accept an inverse-gravity rule just as readily if the code were altered. Because the system lacks the ontic resistance of the physical world to enforce the rule (a thermodynamic consequence for error), the logic remains an arbitrarily imposed constraint, not an intrinsic feature of the system's existence.

In the case of LLMs, the system simulates the statistical profile of human reasoning found in text. When an LLM outputs a causal claim (e.g., “Fire causes smoke”), it is primarily operating by modeling the probability of that token sequence, with no plausible evidence of instantiating the reality of thermodynamic relationship between combustion and particulate matter. The ontic grounding hypothesis suggests that the map, no matter how high-resolution, remains structurally distinct from the territory because it lacks the intrinsic causal powers of the target system. If we accept this premise, treating a text-generation system as a reasoning agent may be a category error akin to treating a weather simulation as a source of hydration.

4.2. The Limits of Multimodal Grounding

A more practical objection comes from recent advances in multimodal AI (e.g., Gemini). Critics might argue that while text-only models are ungrounded, models that ingest video, audio, and robotic sensor data can solve the symbol grounding problem because they connect symbols to sensory inputs.

While multimodality certainly enriches the representational capacity of the system, it does not necessarily solve the fundamental architectural deficit discussed in Section 3. As Harnad (1990) noted, connecting a meaningless symbol (text) to another meaningless symbol (a grid of pixels) is simply “symbol-symbol grounding.” It expands the dictionary to include images, but it does not allow the system to exit the “dictionary-go-round.” To the algorithmic system, a video input is merely a tensor of values, just as text is a vector of tokens. The correlation between the two remains primarily statistical, rather than causal.

This brings us to the risk of biological chauvinism, the observation that the ontic grounding hypothesis would arbitrarily restrict sense-making to carbon-based life. The argument in this paper does not claim that a non-biological system can never possess robust intelligence. It posits that for a system to be semantically grounded, it must plausibly instantiate the structural causal coupling found in living systems. Seth (2026) describes this as mortal computation, which is computation that is inseparable from the material survival of the substrate.

A standard LLM in a data center currently does not face causal consequences. Its errors result in gradient updates, not cessation of function. However, a hypothetical autonomous machine that had to manage its own energy homeostasis to avoid shutdown (system-level death) would begin to approximate the conditions for ontic grounding. The hypothesis is therefore not that silicon cannot think, but that current autoregressive, substrate independent architectures, which are detached from system-level consequences, may lack the feedback loop required for true sense-making. They may remain spectators of reality, optimizing for prediction, rather than participants optimizing for persistence.

5. Addressing the Inconsistent Triad

This paper has argued that the "robustness tension" observed in large-scale AI models may not be an artifact of insufficient training data, but a symptom of a deeper architectural deficit: the ontic grounding gap. By distinguishing between epistemic intelligence (the syntactic manipulation of representations) and ontic intelligence (the sense-making of a grounded agent), we can better understand why current architectures may achieve high fluency while failing at basic logical symmetry and novel generalization.

This architectural diagnosis provides a potential theoretical foundation to begin to address the ethical dilemma posed by Shevlin (2024). Shevlin identifies an "inconsistent triad" in our treatment of AI, consisting of three collectively incompatible premises:

1. **Deep Realism:** Consciousness is a specific natural kind (not merely behavior).
2. **Sentientism:** Moral status depends on the presence of consciousness.
3. **Ethical Behavioural Equivalence (EBE):** Entities that behave indistinguishably from persons should be treated as persons.

The ontic grounding hypothesis offers a principled objection to the third premise (EBE) by challenging the validity of using behavior as an absolute proxy for internal experience. EBE implicitly relies on the assumption that human-like outputs reliably indicate human-like mental

states. However, the persistence of the reversal curse and the ARC-AGI plateau provide empirical support for the claim that behavioral fluency can be decoupled from genuine understanding, which plausibly requires a system-level causal coupling with the environment. Current LLMs represent a class of entities that can be characterized as simulating the outputs of personhood (language, reasoning traces), while seemingly lacking the architectural conditions of personhood (autopoietic grounding). If the simulation of a process (the map) is ontologically distinct from its physical instantiation (the territory), then behavioral equivalence does not necessitate moral equivalence. We could therefore be epistemically justified in withholding moral consideration from systems that lack ontic grounding, regardless of their linguistic performance.

5.1. The Ethical Status of AI

This does not mean AI is without value. Rather, it begins to clarify the nature of that value. LLMs may be viewed not as imperfect agents, but as powerful epistemic instruments, or tools that manipulate the syntactic map of human knowledge with unprecedented speed and scale. Confusing the instrument for an agent (the "unironic mentalising" trap) carries the risk of moral inflation: creating a competitive landscape where the simulated suffering of ungrounded software diverts moral attention and resources from the genuine suffering of sentient beings.

5.2. Toward a Grounded Intelligence

Finally, this analysis could suggest a distinct direction for future research. If the goal of the field remains the creation of genuine AGI, the current paradigm of scaling substrate-independent syntax appears to be progressing in the wrong direction. Progress may instead require a pivot toward mortal computation (Seth, 2026) and neuromorphic architectures, which are systems where the logic of computation is inseparable from the physics of the substrate. Such systems, constrained by genuine thermodynamic risks and homeostatic requirements, might eventually bridge the ontic gap. Until then, however, we must remain clear-eyed about the distinction between the processing of symbols and the grounding of meaning.

References

Allen-Zhu, Zeyuan, and Yuanzhi Li. 2024. "Physics of Language Models: Part 3.3, Knowledge Capacity Scaling Laws." arXiv. https://doi.org/10.48550/arXiv.2404.05405.

Bender, Emily M., and Alexander Koller. 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*, 5185–98. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. FAccT '21. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922.

Berglund, Lukas, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2024. "The Reversal Curse: LLMs Trained on 'A Is B' Fail to Learn 'B Is A'." In *The Twelfth International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2309.12288.

Chalmers, David J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Philosophy of Mind Series. New York: Oxford University Press.

Chollet, François. 2019. "On the Measure of Intelligence." arXiv. https://arxiv.org/abs/1911.01547.

Harnad, Stevan. 1990. "The Symbol Grounding Problem." *Physica D: Nonlinear Phenomena* 42 (1): 335–46. https://doi.org/10.1016/0167-2789(90)90087-6.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. "Scaling Laws for Neural Language Models." arXiv. https://arxiv.org/abs/2001.08361.

Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. "Building Machines That Learn and Think Like People." *Behavioral and Brain Sciences* 40: e253. https://doi.org/10.1017/S0140525X16001837.

Lakoff, George, and Rafael E. Núñez. 2000. *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being*. New York: Basic Books.

Seth, Anil K. 2025. "Conscious Artificial Intelligence and Biological Naturalism." *Behavioral and Brain Sciences* 48: 1–42. https://doi.org/10.1017/S0140525X25000032.

Sharma, Mrinank, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, et al. 2023. "Towards Understanding Sycophancy in Language Models." arXiv. https://arxiv.org/abs/2310.13548

Shevlin, Henry. 2020. "General Intelligence: An Ecumenical Heuristic for Artificial Consciousness Research?" *Journal of Artificial Intelligence and Consciousness* 7 (02): 245–56. https://doi.org/10.1142/S2705078520500149.

Shevlin, Henry. 2021. "Non-human Consciousness and the Specificity Problem: A Modest Theoretical Proposal." *Mind and Language* 36 (2): 297–314.[13] https://doi.org/10.1111/mila.12338.

Shevlin, Henry. 2024. "Consciousness, Machines, and Moral Status." In *Anna's AI Anthology: How to Live with Smart Machines?*, edited by Anna Strasser, 175–196. Berlin: Xenomoi. https://philarchive.org/archive/SHECMA-6.

Shevlin, Henry, Alex Grzankowski, Geoff Keeling, and Winnie Street. 2025. "Deflating Deflationism: A Critical Perspective on Debunking Arguments Against LLM Mentality." arXiv. https://arxiv.org/abs/2506.13403.

Thompson, Evan. 2010. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Belknap Press of Harvard University Press.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, et al. 2022. "Emergent Abilities of Large Language Models." *Transactions on Machine Learning Research*. https://arxiv.org/abs/2206.07682.