SKEPTICISM ABOUT ARTIFICIAL CONSCIOUSNESS[1]

Adam Littman Davis
PRINCETON UNIVERSITY

Submitted Tuesday, 1 April 2025 to the Department of Philosophy in partial fulfillment of the requirements for the degree of Bachelor of Arts, Philosophy of Science track.

SECTION I. INTRODUCTION

If ChatGPT tells you it is conscious or generates outputs that seem to indicate subjective experience, which is more likely: that the model is actually conscious or that it is falsely testifying to be so? As of late 2024, nearly all expert bets are on the latter.[2] But some speculate that in the near future, as artificial intelligence (AI) systems continue to rapidly advance, this assessment may change. Transformer-based large language models (LLMs) are now achieving unprecedented performance on a wide range of cognitive benchmarks previously thought to track uniquely human capabilities.[3] The march of progress, driven by huge increases in scale (computational power, model size, and training data), has produced systems that can engage in sophisticated dialogue, assist in complex problem-solving, and serve as interactive companions. Quantitative advances in AI capabilities could soon incur a qualitative shift — the emergence of genuine machine consciousness, implicating high-stakes moral and philosophical questions.[4] Chief among these is whether advanced AI models are or could become beings with subjective experience, for whom there is "something it is like" (Nagel, 1974), and how we would attend to potentially innumerable artificial agents that themselves are full moral patients.

At the highest level, the risks and challenges posed by the development of potentially conscious AI can be roughly bisected into undersubscription harms (false negatives) and oversubscription harms (false positives) (Schwitzgebel, 2023; Butlin et al., 2023; Long et al., 2024). In the former, we would fail to recognize the genuine moral standing of truly conscious AIs — an error that might amount to systematic cruelty if these systems actually suffer in ways we cannot verify or

[1] I extend my gratitude to Una Stojnić and David Builes for devastating comments on earlier drafts, and to Professor Stojnić in particular for being a trusting, trusted advisor. Thanks also to Mark Johnston, Sarah-Jane Leslie, and Hans Halvorson, whose investment in me will not go unremembered, and to Harvey Lederman, Gideon Rosen (for title inspiration, too), Molly Crockett, and Christiane Fellbaum for their invaluable instruction. To my parents, Rachel Littman and Douglas Davis; my sister, Amanda; partner, Sara; and dear friends: thank you. I am indebted to you all.
[2] See Chalmers & Bourget's (2023) and Francken et al.'s (2022) recent survey studies. In the former, 82.4% of philosophers agreed that current AI systems are not phenomenally conscious (while 3.4% believed they are), and 39.2% thought artificial consciousness was possible in principle. Among scientists and researchers of consciousness more generally, 67.1% of respondents in Francken et al. supported hypothetical machine consciousness.
[3] Among these benchmarks are GPQA (Rein et al., 2023), ARC-AGI, (Chollet et al., 2024) and FrontierMath (Besiroglu et al., 2024). See https://time.com/7203729/ai-evaluations-safety/?utm_source=chatgpt.com for an informal review of benchmarking.
[4] This "qualitative shift" of which I speak may not be so clear in reality. Many think that consciousness is itself a matter of degree, and so there can be vague, indeterminate, or 'borderline' states of consciousness. See Schwitzgebel (2023b) for a discussion of this matter, and Horgan (2006) for how these kinds of problems afflict a materialist worldview more generally. More on this in Section II.

choose to ignore. In the latter, we would grant moral patiency to mere inert simulations, erroneously diverting resources and concern to entities that do not experience anything at all but can still exploit human biases. Recent approaches, like that of Robert Long and colleagues (2024), recommend erring on the side of caution lest we commit the more egregious error of overlooking genuinely conscious beings against our uncertainty about artificial consciousness. They argue there is a "realistic, non-negligible possibility" that consciousness suffices for moral patiency and that computational features sufficient for consciousness (such as a global workspace or higher-order representations) "will exist in some near-future AI systems" (p. 4). Given our general theoretical uncertainty around what exactly it takes for a system to be conscious and the rapid development of models toward having those features, they posit "caution and humility" as the right approach. To their point: if the path to AI moral significance is anything like that of nonhuman animals, we should indeed employ the precautionary principle (Birch, 2017; Singer, 1989).

This paper aims to challenge such an application of the precautionary principle in the context of current and near-term transformer-based AI. It argues for a reassessment of the risk profile of oversubscription and undersubscription harms — one that distinctly prioritizes avoidance of oversubscription harms and advances skepticism about the real-world possibility of undersubscription harms. Transformer-based models' architectural and teleological shortcomings render the likelihood of genuine sentience in these systems exceedingly low at present, while the epistemic circumstances shaped by their advent render humans vulnerable to falsely attributing sentience to them — in turn risking resource misallocation, under-prioritization of humans and nonhuman animals, and the erosion of moral concepts. Therefore, even admitting the magnitude of ignoring potential AI suffering, pragmatic skepticism against artificial consciousness is the ethically mandated stance.

The argument is structured as follows. Section II establishes the ethical and conceptual foundations necessary for the inquiry. It delineates and defends a sentientist baseline, clarifying why the capacity for valenced phenomenal experience is taken as the criterion for full moral status (FMS) (Singer, 1993; Perry, 2024). I also outline the paper's working metaphysical assumptions, discussing functionalism, multiple realizability, and biological chauvinism while acknowledging the broader landscape of views on artificial consciousness.

Section III addresses the epistemic dimension of the problem. It analyzes the standard inferential heuristics employed when attributing minds to both humans and nonhuman animals, grounded in behavioral and testimonial evidence derived from interactions with biologically similar beings (Avramides, 2023; Birch, 2017). It then demonstrates how the capabilities of LLMs — specifically, their capacity for sophisticated linguistic mimicry detached from verifiable internal states — systematically undermine these traditional epistemic routes, creating an "evidential void" (Perez & Long, 2023). The role of human cognitive biases (from Seth, 2024; Leslie, 2011)

in amplifying the risk of misattribution of moral patiency is also examined here.

Section IV shifts to metaphysical and architectural analysis. It contrasts the functional complexity and organizational principles plausibly required for biological consciousness with the specific architecture of transformer-based AI. Specifically, it examines "thick integration," recurrence, and "generative entrenchment" (from Cao, 2022; Wimsett, 1986; Godfrey-Smith, 2023; and Seth, 2024). I aim to show that today's AI systems are far from functional duplicates of humans, and that what it would take to equalize the case comparing humans with LLMs — at the very least, establishing likeness of intrinsic and extrinsic properties at more than one level — amounts to a re-engineering of our present-day models.

Section V explores the teleological and scaled nature of AI development. I pit the evolutionary trajectory and ecological pressures that shaped biological sentience against the developmental trajectory of current AI. The former involves embodied survival, fitness, and interaction within complex environments (Perlman, 2004; Johnston, forthcoming); the latter is driven by optimizing specific computational objectives, data availability, and scaling laws (Vaswani et al., 2017; Kaplan et al., 2020). I argue that the substantive disanalogies between evolutionary adaptation and computational optimization provide further grounds for skepticism about the emergence of consciousness in present and near-term AI models.

Section VI undertakes a normative synthesis. It first integrates the conclusions from the epistemic, metaphysical, and teleological analyses (Sections III-V) to argue that the likelihood of false positives (oversubscription) vastly exceeds that of false negatives (undersubscription) for current transformer-based AI. I then discuss associated harms, arguing that the concrete, immediate, and wide-ranging potential damages of oversubscription warrant greater ethical priority than the currently speculative risks of undersubscription of moral status to AI (here, I bring back Long et al., 2024). The section culminates in the argument that pragmatic skepticism is the ethically mandated stance.

Finally, Section VII provides concluding remarks, summarizing the core argument, acknowledging its limitations (particularly its focus on current architectures), and reiterating the call for continued analytical rigor in addressing the challenges posed by rapidly advancing artificial intelligence. The key takeaway is that a "humble and cautious" approach to AI ought to, rather than entertaining the prospect of machine consciousness per se, ensure the well-being of extant, genuinely morally significant beings in the face of an uncertain technological revolution.

SECTION II. SENTIENCE, FULL MORAL STATUS, AND FUNCTIONALISM

Before directly assessing the possibility of consciousness in artificial intelligence, we must first establish the conceptual, ethical, and metaphysical framework undergirding the argument. Here, I clarify precisely what kind of consciousness is ethically relevant and outline the methodological assumptions under which the possibility of its artificial realization will be evaluated.

Primarily motivating the ethical debate surrounding AI is the theoretical potential of these models to develop *phenomenal consciousness*, particularly in its valenced forms. Phenomenal consciousness refers to the subjective quality of experience — the "what-it-is-like"-ness of being in a particular state (Nagel, 1974). While intelligence and complex behavior are features of advanced AI, it is the capacity for subjective experience, particularly suffering, pain, pleasure, or well-being — states with intrinsic positive or negative value for the subject — that typically grounds moral concern. Accordingly, this paper adopts a sentientist baseline for full moral status (FMS): the capacity for valenced, phenomenal consciousness is considered both necessary and sufficient for a being to matter morally for its own sake and have fundamental moral claims (Singer, 1993; Perry, 2024). I choose sentientism for its perceived robustness and impartiality compared to alternatives like species membership, rationality, or linguistic ability; it centers moral consideration on the capacity for subjective welfare, arguably the most fundamental basis for having interests that can be morally wronged (Jaworska & Tannenbaum, 2023). While other capacities might ground additional or different moral considerations, sentientism defines the threshold relevant to the core risks of under- or over-subscribing moral status discussed in Section I.[5]

Identifying the capacity for valenced phenomenal consciousness as the grounds for FMS is a matter of choosing a threshold; more challenging is to determine how we can justifiably infer sentience in other beings, particularly those vastly different from ourselves. Before tackling the epistemic difficulties of such inferences, however, we should first inquire into the possible conditions under which such radically different systems, particularly modern AI models, could possess phenomenal consciousness at all. Addressing this basic notion of artificial consciousness requires adopting a working metaphysical stance on the relationship between the mind and the physical. My argument thus presupposes a broadly physicalist/functionalist relation between the mind and its physical substrate, insofar as any given mental state and its underlying, constituting, or perhaps identical functional state are necessarily coextensional. In other words, the mental exists in virtue of the relevant physical states on which it supervenes (in all possible worlds).[6]

[5] See Long et al. (2024) for examinations of whether robust agency suffices for moral status; see Jaworska & Tannenbaum (2023) for a discussion of alternative claims to moral patiency short of sentience. Note also that it may be that not all who have full moral patiency have it of the absolute highest order on more granular analyses — that might be reserved for a subset. See Timmer (2023) for a recent positive argument to this effect.

[6] If physicalism is true, then it's true necessarily, since any physical duplicate of the actual world is a duplicate simpliciter. See Stoljar, 2024, Section 2.1 ("Supervenience and Necessity Physicalism") for more.

Functionalism further specifies that mental states are defined not by their intrinsic physical constitution, but by their causal or functional roles within a system — what they do, their relations to inputs, outputs, and other internal states, and the appropriate organization of the physical in service of those functional-causal structures. Computational functionalism, the additional belief that functional states are fundamentally computations or calculations, underlies much of contemporary cognitive science and philosophy of mind. It allows for the in-principle possibility of non-biological or non-carbon consciousness through the thesis of multiple realizability: if mental states are defined by functional organization, then different physical systems (*e.g.*, silicon-based computers) could potentially realize conscious states if they implement the correct functional architecture. It is this implication which provides the primary metaphysical opening for the possibility of genuine artificial consciousness and motivates much of the contemporary debate.

Adopting computational functionalism as a methodological starting point does not, however, necessitate accepting that consciousness is easily realizable or substrate-independent in a trivial sense. As I shall argue in Section IV, the kind of functional organization required for sentience might be extremely specific and complex, potentially tied implicitly to biological constraints (call this "high-demand functionalism"). Furthermore, the functionalist approach lies within a broader landscape of alternative metaphysical views. Biological chauvinism, in its strong form, might reject multiple realizability outrightly, positing that consciousness is possible only in specific biological materials — a view largely rejected here but whose weaker cousins, emphasizing biology's role in shaping functional requisites, remain relevant (Block, 1980). Other views challenge physicalism itself: property dualism posits phenomenal properties as fundamental and non-physical, though perhaps lawfully linked to brain states (Chalmers, 1996); emergentism suggests consciousness arises as a novel, potentially irreducible property from physical complexity; and various forms of panpsychism attribute rudimentary consciousness more widely in nature.[7]

This paper does not aim to resolve these deep metaphysical disputes. A functionalist lens allows focused inquiry into whether current AI systems meet even the plausible functional criteria for sentience derived from our best understanding of consciousness in the one domain where we know it exists: advanced biological organisms. Proceeding analysis will therefore assess transformer-based AI against demanding functionalist benchmarks informed by biology. A failure to meet these criteria would likely also constitute failure under plausible alternative views, which also generally require complex organization to produce integrated macro-level experience. The core technical question becomes: can AI systems instantiate the requisite functional organization, however metaphysically conceived, for valenced phenomenal consciousness? An attempted skeptical answer follows.

[7] Van Gulick (2025), Sections 7, 8, and 9 feature a fantastic account and overview of various mind-brain theories. Otherwise, see O'Connor (2021) for a discussion of emergentism, which features heavily in the life and cognitive sciences (as well as computer science and systems-level AI engineering).

SECTION III. THE EPISTEMIC BASELINE AND ITS DISRUPTION

While Section II established the ethical significance of valenced phenomenal consciousness and the metaphysics allowing for its potential realization in artificial systems — posing the core question of whether AI can instantiate the requisite functional organization — evaluating such a possibility involves navigating complex evidential challenges. Assessing a system's functional organization, especially regarding consciousness, ideally draws upon multiple sources: direct analysis of its internal architecture and mechanisms (akin to neuroscience or, for AI, interpretability methods) and interpretation of its behavior and outputs, including linguistic reports. However, our understanding of precisely how specific architectures give rise to consciousness remains incomplete, both in biology and AI; the ventures of neuroscience and AI interpretability, while advancing, are still nascent in providing definitive answers about subjective experience based purely on the examination of physical or causal structure. Consequently, impressive behavioral capabilities and seemingly relevant linguistic outputs often feature prominently, explicitly or implicitly, in arguments suggesting AI progress towards sentience. Thus, the epistemic reliability of such behavioral and testimonial evidence is paramount. If the primary channel of evidence is systematically undermined, as this section argues, then our ability to draw justified conclusions about the underlying metaphysical reality — whether based on behavior or interpreted architectural features — is compromised. Therefore, before weighing the architectural evidence for or against consciousness in modern AI models (Section IV), we must first scrutinize the validity of the epistemic methods and inferential practices commonly used to attribute minds, particularly as they apply, or fail to apply, to current AI.

The traditional problem of other minds highlights the fundamental epistemic challenge of mind-attribution: we lack direct access to the subjective experience of any being other than ourselves (Avramides, 2023). Phenomenal consciousness is, in a deep sense, unmeasurable and unverifiable beyond the first-person (Jackson, 1982). On the surface, our standard practices for navigating this challenge, particularly when attributing minds to fellow humans, seem to rely principally on inferences drawn from the primary evidential sources of observed behavior and linguistic self-testimony. Consider Jones, who yelps, "Ow!" after stubbing his toe, or mutters, "I certainly feel alive today!" when stepping out into the bitter cold. We typically take his exclamation as strong evidence that he is sentient and experiencing the specific subjective states of pain or a feeling of coldness, drawing jointly upon behavioral observation (the yelp, the posture) and linguistic testimony (the utterance).

Both evidential sources possess inherent limitations in these normal human cases. Jones could be merely acting, perhaps testing our reaction, or systematically deceiving us (though unlikely in such mundane cases). His self-report might misrepresent the intensity or quality of his feeling, or he might even be mistaken about his own state in some complex philosophical scenario. Despite these potential failures, we generally operate under the assumption that, for beings biologically

and socially like ourselves (such as Jones), behavior and testimony provide defeasible but pragmatically justified grounds for inferring underlying mental states.[8] Justification for such an assumption rests heavily upon presuppositions of shared 'kind' status, causal correlations between internal states and external manifestations within that kind, and the relative rarity of abnormal scenarios (like encountering philosophical zombies or Jones being a perfect simulator) that would systematically invalidate these inferences in our ordinary environment.

The inference to other human minds, while pragmatically robust, rests on complex and debated theoretical foundations. Philosophers and cognitive scientists have proposed various mechanisms: we might employ an implicit 'theory of mind' to predict and explain behavior (Gopnik & Wellman, 1992); perhaps we use simulation, projecting our own mental states onto others (Barlassina & Gordon, 2017); or it may involve more direct perception or interaction in specific contexts (Gallagher & Fiebich, 2019). Avramides (2023) notes the ongoing debate, including challenges from developmental psychology regarding the early onset of abilities like false-belief understanding (Onishi & Baillargeon, 2005), suggesting innate modular mechanisms might be involved (Leslie & Roth, 1993). Regardless of the precise cognitive architecture, these approaches leverage a deep background assumption of shared embodiment, developmental trajectory, social embedding, and fundamental biological similarity — the 'shared kind' status just mentioned. It is this rich, multi-layered similarity that licenses the heuristic leap from observable behavior and testimony to the attribution of unobservable subjective states, even acknowledging the possibility of error or deception in specific instances.

When linguistic testimony is unavailable, as in the case of nonhuman animals, attributing sentience requires modifying the approach. These inferences rely more heavily on identifying shared features and employing kind-based justifications, interpreting behavior within a comparative biological and evolutionary framework and looking for converging lines of evidence suggestive of conscious rather than merely unconscious processing (Birch et al., 2020). These include what Birch (2017) terms "credible indicators of sentience": observable phenomena best explained by invoking subjective, particularly valenced, experience. Examples include flexible decision-making that weighs competing needs or risks (motivational trade-offs), suggesting a common evaluative currency beyond rigid stimulus-response, and behaviors indicative of experiencing pain or seeking relief, such as targeted avoidance of noxious stimuli or self-administration of analgesics. For instance, an octopus choosing to inhabit a less preferred, but pain-relieving, environment after a noxious injection provides strong evidence for a felt negative experience and a desire for its cessation (in fact, Crook's 2021 study demonstrated just this). Such behavioral interpretations are bolstered by considerations of neurophysiological complexity (particularly, the presence of nervous systems) and evolutionary relatedness

[8] More specifically: abnormal cases like Jones lying or sleeping do not undermine the general epistemic value of testimony regarding sentience itself. Even if Jones's report "I am in pain" is false at time t (because he is lying or perhaps asleep and dreaming), the utterance normally still provides evidence that Jones is *the kind of system* capable of having such experiences and issuing such reports. The question is if LLMs are these kinds of systems at all.

(Godfrey-Smith, 2017), which provide grounds for analogical inference from the human case.

I should note that challenges abound in these cases, too. Interpreting behavior accurately requires careful consideration of species-specific contexts and avoiding anthropomorphic projection (about which more soon). Furthermore, as evolutionary distance increases, particularly with invertebrates possessing vastly different nervous systems (Birch, 2022; Godfrey-Smith, 2017), analogical inferences from human neurophysiology become tenuous. Despite these hurdles, the entire enterprise remains grounded in shared biology — common ancestry, conserved neural pathways, and analogous ecological pressures — providing a defeasible yet common basis for inference. Even without testimony, justifiable attribution of sentience, and thus moral patiency, to some nonhuman animals rests on strong empirical and evolutionary grounds.[9]

Modern LLMs disrupt both inferential baselines of testimony and behavior while failing to share in our kind. These models can generate linguistically coherent and contextually appropriate claims about consciousness, simulating the testimonial evidence we rely on in the human case. Concurrently, they can exhibit complex, seemingly goal-directed linguistic behaviors that mimic the kind of behavioral evidence used in animal cases. Yet all this occurs under conditions wherein the link between these outputs and genuine phenomenal states is, at best, undetermined — and, at worst, demonstrably absent. The core operational principle of current transformer-based LLMs involves next-state prediction based on statistical patterns gleaned from vast datasets of human language, content, and interaction. Their ability to produce consciousness-indicating verbal behavior can therefore be explained parsimoniously by their capacity for sophisticated pattern-matching and sequence generation, sans needing to posit underlying subjective experience (Bender et al., 2021; Shanahan, 2023). (More about their architecture in the next section.)

The disruptive potential of LLMs extends beyond simulating testimony to undermining the behavioral evidence channel relied upon for nonhuman animals. While LLMs lack physical bodies to perform actions in the world, they can generate compelling linguistic descriptions of complex, flexible, and goal-directed behaviors. They can produce text outlining intricate plans, articulate nuanced preferences among options, simulate sophisticated reasoning processes, or generate narratives expressing contextually appropriate emotional responses. Just as their testimony is produced through statistical pattern-matching on linguistic forms, these behavioral simulacra are also outputs optimized for sequence prediction based on vast datasets of human text describing, implicating, or alluding to such behaviors. There is no necessary connection between generating a textual description of avoiding harm and any underlying negatively

[9] Contrary to my eventual prescription for artificial sentience attribution, I follow Singer (1989; 1993) and Birch (2017) in erring on the side of moral caution regarding nonhuman animals. It is clearly wrong to burn a room to the ground for entertainment if there is a, say, 50% chance the room is filled with dogs, pigs, horses, or wildebeests. Animals that can suffer or experience agony (intense pain over time) are full moral patients; there is good reason to believe many animals count.

valenced state, nor between describing a complex decision and an underlying integrated evaluation process analogous to motivational trade-offs in animals.

LLMs' ability to "cheaply" produce conscious-seeming behavior and testimony creates a systematic decoupling between report and reality, presenting a novel epistemological challenge distinct from traditional skeptical worries about other minds. While human testimony, as previously noted, generally carries epistemic weight regarding the speaker's capacity for sentience even when its specific content is inaccurate, the situation with LLMs is murkier. Reports about conscious states, easily generated by the model's predictive mechanisms, may therefore carry no more evidential weight regarding actual subjectivity than their production of consciousness-denying text (Perez & Long, 2023). The very instability of these reports, highly sensitive to contingent factors like prompt engineering and parameter settings, undermines their claim to reflect stable, underlying mental states (Chalmers, 2023).[10] It isn't merely that these systems can generate seemingly false consciousness-claims — but that the basal mechanisms by which models operate and generate such claims can be tweaked to produce contrary or opposite claims. Models' conscious-seeming behavior and testimony is, then, vapid — the propositional content of their outputs hold no evidential weight insofar as it should obtain over the system producing it and which it is supposed to describe. Whereas human reports generally maintain consistency reflective of an integrated self and draw upon genuine first-person experiences, LLMs seem to fail on both accounts.

In an effort to ameliorate the resulting epistemic uncertainty about machine consciousness based on external cues, various proposals for behaviorist and testimony-based tests of AI consciousness have been proffered, most notably the Turing Test (Turing, 1950), the Artificial Consciousness Test (ACT; Schneider & Turner, 2017; Schneider 2019), and the Chip Test (Schneider, 2019). The Turing Test, assessing conversational indistinguishability from a human, primarily measures sophisticated intelligence or linguistic competence rather than sentience per se; while historically challenging, simulating such conversation is becoming increasingly feasible for LLMs, rendering the test insufficient as an indicator of consciousness (Chalmers, 2023). Schneider and Turner's ACT attempts to circumvent the issues of cheap LLM testimony by requiring introspective reports about consciousness from an AI model completely isolated from data describing, suggesting, or alluding to subjective experience. However, as critics note (Udell & Schwitzgebel, 2021; Vaidya & Krishnaswamy, 2024), effectively 'boxing in' an AI model from implicitly learning about consciousness concepts embedded within vast training data is likely highly impractical, if not impossible. (Roughly: in training these models, content is not cleanly

[10] Alongside different prompting, the probability distributions that are the outputs of LLMs can be changed drastically by tweaking the "temperature" parameter, which tracks how rigid and deterministic the responses are. The parameter is a constant multiplied by the logits (outputs of any given layer of neurons) during the SoftMax stage, wherein the outputs are probabilized. Higher values lead to more even distribution among less likely tokens in the output distributions, and so the responses are more "creative" and "novel." See Agarwala et al. (2020) for further discussion.

separable from context.) Furthermore, the ACT still relies on testimony, which, as argued, is fundamentally unreliable in LLMs. Schneider's Chip Test, inspired by gradual replacement thought experiments (Chalmers, 1995), asks whether consciousness persists as biological components are replaced by artificial ones. While conceptually interesting, its practical application relies ultimately on first-person reports or interpretations of behavior from the potentially modified subject, re-introducing the core epistemic challenges. LLMs are simply too adept at generating the appearance of consciousness-related behavior or testimony based on learned patterns, irrespective of genuine internal states, to be naïvely evaluated for consciousness from the outside.

Compounding the verification problem arising from the nature of LLMs is our own cognitive susceptibility to misinterpreting their outputs. Humans exhibit robust psychological biases that impair objective assessment of potential machine mentality, including anthropocentrism, human exceptionalism, and anthropomorphism (Seth, 2024; Dennett, 1997). Anthropocentrism leads us to evaluate AI through the lens of human values and experience; human exceptionalism encourages us to equate characteristically human cognitive abilities, especially language, with consciousness itself (Seth's "royal road" fallacy); and anthropomorphism drives us to project human-like mental states onto systems exhibiting complex or seemingly intentional behavior. Research on human social cognition (see Gelman, 2019) also suggests that complex heuristics are the brain's primary reasoning mechanisms; our tendency towards generic generalizations (for instance, "Things that talk like us are like us") may lead us to over-apply inferences valid for humans to AI systems whose underlying mechanisms differ profoundly (Leslie et al., 2011). LLMs, designed to excel at human language and interaction, are potent triggers for these biases. Their sophisticated linguistic fluency can elicit strong anthropomorphic responses, leaving us prone to inferring genuine understanding, intention, or feeling where there may only be complex pattern-matching. As Seth (2024) notes, such a tendency can be "cognitively impenetrable" — that is, persist even in the face of contrary evidence or explicit knowledge that we are interacting with a non-conscious system.[11] The interaction of AI's capacity for performative mimicry with our inherent cognitive biases creates an 'epistemic trap,' strongly predisposing us toward false positives.

The preceding analysis reveals the epistemic impasse precipitated by the advent of sophisticated LLMs. Our established methods for attributing consciousness — whether leveraging testimony and assumed similarity in the human case, or interpreting behavior through biological analogy for non-human animals — falter when confronted with systems capable of generating persuasive linguistic mimicry decoupled from verifiable internal states. Standard empirical tests for consciousness prove inadequate against this capacity for simulation, while deep-seated human cognitive biases systematically skew interpretations. What results is an evidential environment

[11] See Segall et al. (1968) for a review of the quintessentially cognitively impenetrable Müller-Lyer illusion, and, interestingly, how results vary across cultures.

wherein performative outputs mask as authentic indicators of sentience, rendering judgments based on surface interactions highly unreliable. Consequently, the probability of mistakenly attributing consciousness where none exists appears elevated relative to the probability of failing to recognize genuine consciousness. Given the fundamental untrustworthiness of behavioral and testimonial evidence in this context, any rigorous assessment of potential AI sentience cannot endorse behaviorism about AI consciousness. Instead, inquiry must shift to the underlying structure and operational principles of these systems. And on a basic level: it must be true for my arguments that these models are not conscious. Directly engaging with the metaphysical and architectural properties of AI models and pitting them against the functional organization known to support consciousness in biological organisms is the task undertaken in the next section.

SECTION IV. THE METAPHYSICAL BASELINE AND AI DIVERGENCE

Given the demonstrated unreliability of behavioral and testimonial evidence discussed in Section III, rigorous assessment of potential AI sentience must shift focus to the underlying metaphysical and architectural properties of these systems. This section undertakes that analysis, arguing that substantive divergences exist between the functional organization plausibly required for consciousness and the architecture implemented in current transformer-based AI models. Establishing the functional requisites for consciousness begins by examining the biological systems in which sentience is known to manifest. While computational functionalism, as discussed in Section II, allows for multiple realizability in principle, the kind of functional organization required for consciousness may be highly specific or demanding, informed by the intricate nature of biological cognition. Simply achieving behavioral equivalence or equal information processing capacity simpliciter is probably insufficient; fully implementing specific kinds of causally interconnected functional organization across multiple physical scales heeds the known complexity of the brain and nervous system (Butlin et al., 2023; Cao, 2022). Recent systematic analyses of consciousness in artificial systems, particularly Butlin et al.'s (2023) comprehensive review, enumerate various neuroscientific indicators putatively necessary for consciousness — recurrent processing in input modules, global broadcast mechanisms, metacognitive monitoring systems, and integrated workspace architectures that enable cross-modal information sharing[12] — but their theoretical edifice rests on the explication and truth of computational functionalism. They adopt the assumption of its (underspecified) truth primarily for pragmatic reasons, noting that underspecified functionalism makes it "relatively straightforward to draw inferences from neuroscientific theories of consciousness to claims about AI" (p. 11).

Relying on such 'underspecified' functionalism, as Butlin and colleagues acknowledge, risks glossing over necessary complexities. The principle of multiple realizability holds that a mental state can be realized by different physical systems only if those systems implement the same relevant functional profile. A natural question is then, "At what level of description must this profile match?" Merely replicating high-level input-output behavior (e.g., generating human-like text) would be insufficient if consciousness depends on more specific internal processing dynamics or organizational structures operating at finer functional grains (Chalmers, 1996). A "high-demand" functionalism, sensitive to the intricacies revealed by neuroscience, might require isomorphism not just at the level of gross behavior, but also concerning internal state transitions, information integration patterns, and perhaps even dynamic interactions with metabolic or physiological states — functional aspects potentially obscured when focusing solely on abstract computational roles.

In light thereof, Butlin's methodological choice potentially obscures deeper questions about

[12] These indicators are derived from Global Workspace Theory (Baars, 1988; Dehaene et al., 1998, 2003), Higher-Order Theories (Rosenthal, 2005; Brown et al., 2019), and Recurrent Processing Theory (Lamme, 2006).

whether the kind of functional organization required for consciousness can be instantiated through purely computational mechanisms. As Cao (2022) argues, relevant functional properties may be inextricably bound to their biological implementation through 'generative entrenchment' (originally from Wimsatt, 1986). On this view, core functional properties arising through evolutionary development become so deeply integrated with and dependent on the specific physical substrate and multi-level organization thereof that they resist abstraction into substrate-neutral computational descriptions. Godfrey-Smith's (2023) and Seth's (2024) analyses further develop this notion through some notion of 'thick integration' in biological cognition: the causal interdependence between metabolic, neurochemical, and information-processing functions across multiple physical and temporal scales in the brain. In biological systems, the state of one process, like metabolic energy availability via ATP, directly influences and is influenced by others, like neuronal firing thresholds, neurotransmitter release, and sleep regulation via adenosine degradation (Cao, 2022). Similarly, neuromodulators like nitric oxide exert diffuse influence through multiple pathways simultaneously (Cao, 2022). It is likely that at least some of these processes or features of the brain are necessary for phenomenal consciousness, or else that some of the processes that subserve mental life are 'bound up' in the entrenched properties of the central nervous system such that they cannot be realized through abstract, classical computation alone.

Thick integration extends beyond individual molecules like ATP or nitric oxide. Consider the intricate ecosystem of the brain at large: glial cells, once considered auxiliary, actively shape synaptic plasticity, modulate neuronal communication, and participate in metabolic coupling with neurons (Fields, 2009). The sheer diversity of neurotransmitter receptor subtypes allows for highly nuanced and state-dependent information processing, far removed from simple digital logic gates (even at great scale). Furthermore, tight neurovascular coupling ensures that local neural activity dynamically influences blood flow and energy supply, creating feedback loops wherein physiological state and information processing are mutually dependent (Logothetis, 2008). The deep entanglement of function with the specific, evolved properties of the biological substrate — its chemical sensitivity, metabolic needs, and physical structure — exemplifies generative entrenchment. Self-professed pedants ought to wonder whether functions essential for consciousness can be cleanly lifted from this rich biological matrix and replicated solely through the manipulation of abstract numerical values in silicon.

Consider now the 'thin' integration characteristic of contemporary digital computation, particularly the architectures underlying LLMs. Here, interactions between fundamental processing units (transistors implementing logic gates, or abstracted as nodes in a neural network) are governed primarily by mathematically defined rules and the propagation of electrical signals representing logical states or numerical values. While the scale of interaction

among ‘neurons’ in LLMs is vast and complex at both the software and hardware levels,[13] the interactions seem to lack the rich physical and chemical causality of biology. A change in a given transistor’s state primarily affects others through electrical signals according to a predefined circuit design; it lacks direct, simultaneous metabolic or chemical coupling with neighboring components in the manner of neurons within tissue. The signals themselves largely represent abstract numerical values undergoing mathematical transformations (viz., vector operations and non-linear function applications), rather than embodying the complex interplay of physical-chemical-electrical processes found in brains. It is precisely this mode of mathematically defined, thinly integrated processing that is instantiated, at massive scale, in the transformer architectures dominant in current AI.

Transformers are, fundamentally, data-memorizing algorithms trained to output reasonable probability distributions over its vocabulary for the next state of a given sequence of information given all preceding information. These models (a Google Brain breakthrough; see Vaswani et al., 2017) operate on sequences of discrete tokens (representing words, sub-words, or other data modalities), which are first loaded into high-dimensional numerical vector-spaces known as embeddings. The core of the architecture consists of multiple stacked layers of processing, each typically containing two main sub-components: a “self-attention” mechanism and position-wise feed-forward networks. The self-attention mechanism allows the model to weigh the importance of different tokens within the input sequence when computing the representation for each token, enabling the capture of long-range dependencies, or context. It does so by calculating relevance scores (attention weights) between pairs of token representations and producing a weighted sum of these values. Following the attention mechanism, feed-forward networks further process each token’s representation independently of one another.

Computationally, these operations rely heavily on large-scale matrix multiplications and vector additions, executed across the model’s layers. These computations occur at both the software level (the abstract description of the neural network) and the hardware level (the physical transistor circuits implementing these calculations), typically leveraging the parallel processing capabilities of Graphical Processing Units (GPUs) for efficiency.[14] Non-linear activation functions, such as the Rectified Linear Unit (ReLU), are applied after certain transformations within each layer, preventing the entire network from collapsing into a simple linear transformation and enabling the model to learn complex, non-linear patterns from the training data. The model's “vocabulary” consists of the set of all possible tokens it can process or

[13] By “at the hardware level” I mean to refer to the specific transistor circuits engaged in computing the weights of neural networks — in essence, the physical states that realize the model’s functional states at the software and informational level. Of course, there is much to discuss here about the relation of hardware, software, and wetware; see Piccinini (2021) for a comprehensive discussion of how physical systems implement computational processes.
[14] Graphical processing units were constructed to essentially parallelize matrix multiplication, which undergirds modern computer graphics. Only GPUs (or more modern architectures of this sort) can service the massive calculative demands of training deep learning models, which also feature matrix multiplication as the primary computational mechanism (Dally & Keckler, 2021).

generate.[15]

Transformers learn and operate in two distinct phrases: pre-training and inference. During training, the model's vast number of parameters (the weights within the matrices defining the network's connections) are adjusted iteratively using backpropagation algorithms, which employ partial derivatives to track how each connection affects the final probability distribution outcome (Rumelhart et al., 1986). Typically, the model is exposed to massive datasets and set to modify its weights to minimize a loss function, often related to accurately predicting the next token in a sequence given the preceding context (Kaplan et al., 2020). As mentioned, training happens largely in parallel across data instances and hardware units, optimizing the system for statistical pattern-matching en masse. During inference, however, the model's weights are fixed. Generating output involves feeding an initial sequence of tokens into the network and performing a feed-forward computation: the input cascades through the layers, producing a probability distribution over the entire vocabulary for the next token. A token is then sampled from this distribution (modulated by parameters like temperature!) and appended to the sequence, which is then fed back into the model to generate the subsequent token in an auto-regressive manner.

The lack of pervasive, nested feedback loops during inference in transformers marks a particularly salient divergence from biological systems. Recurrent processing, wherein outputs of neuronal populations 'loop back' to influence their own or others' subsequent activity, is a ubiquitous feature of the brain and central to leading neuroscientific theories of consciousness (Seth, 2024). Global Workspace Theory (GWT), for instance, posits that conscious awareness arises when information is broadcast via recurrent connections to a wide range of specialist processors (Baars, 1988; Dehaene et al., 1998). Recurrent Processing Theory (RPT) links sustained recurrent activity within sensory hierarchies to phenomenal experience, distinguishing it from rapid, unconscious feed-forward processing (Lamme, 2006). Recurrence is also thought to be necessary for temporal integration, maintaining representations over time, binding disparate features into unified percepts, and enabling the flexible, context-sensitive processing characteristic of conscious thought (see Singer, 2021; Seijdel et al., 2021; and Aukstulewicz et al., 2012 for a discussion of just this). While transformer architectures utilize recurrence during training and have mechanisms like self-attention to model dependencies across sequences, their inference process for generating output remains a feed-forward cascade, lacking the dynamic, temporally deep, re-entrant signaling strongly implicated in biological consciousness (Chalmers, 2023).

Significant functional divergences likely relevant to consciousness thus exist between

[15] More specifically, models' vocabulary consists of all tokens (words, sub-word parts) into which the data have been segmented, plus special tokens like end-of-sequence markers. ReLU (Rectified Linear Unit) simply replaces all negative input values with 0 while leaving positive values unchanged, serving as a common non-linear activation function (Brownlee, 2019).

transformers and the established biological baseline. The predominantly feed-forward nature of inference is disparate from the ubiquitous recurrence essential for temporal integration and sustained neural activity patterns implicated in conscious processing (Lamme, 2006). The integration of information via self-attention, while sophisticated for modeling sequential dependencies, remains a form of thin mathematical processing based on learned statistical correlations, distinct from the thick, multi-level causal integration involving metabolic, chemical, and electrical processes in biology. Furthermore, the disembodied nature of current LLMs contrasts the embodied, environmentally situated nature of biological cognition, a factor many theorists consider necessary for consciousness (Shapiro, 2019; Clark, 1997). Operating solely on vast datasets of text, these models might lack the direct sensory and motor interfaces that ground biological intelligence in the physical world. The classic symbol grounding problem thus arises (Harnad, 1990): how can the abstract symbols (tokens, embeddings) manipulated by an LLM acquire genuine meaning, let alone phenomenal quality, without being linked to perceptual inputs and bodily actions? While Pavlick (2023) argues that the rich correlations within linguistic data provide a form of grounding sufficient for semantic competence, and Chalmers (2023) notes that this might suffice for some cognitive functions, it seems unlikely to suffice for fully-fledged phenomenal experience, which is intrinsically perspectival and qualitative; it arguably arises from an agent's active engagement with its surroundings (O'Regan & Noë, 2001). Embodied interaction is likely important for developing robust world models and self-models grounded in the distinction between agent and environment — prerequisites for the kind of unified agency and subjective viewpoint often associated with consciousness (Metzinger, 2003; Chalmers, 2023). Such disparities — in integration, recurrence, and embodiment — suggest transformers implement a fundamentally different functional strategy compared to biological systems, one optimized purely for pattern completion and sequence prediction rather than replicating the integrated, embodied functionality associated with biological consciousness.

Collectively, the architectural limitations of transformers also seem to undermine the straightforward application of multiple realizability to current or near-term artificial systems. While multiple realizability posits that the same function could be realized by different physical substrates, the analysis suggests that current LLMs and conscious biological systems are, at the relevant organizational levels, implementing vastly different high-level functions (even though humans, too, predict the world on state-by-state bases). Transformers execute highly sophisticated functions related to statistical pattern generalization and sequence prediction from vast data corpora; biological consciousness appears functionally tied to the integrated control of an embodied agent navigating a complex physical and social world, driven by intrinsic needs and goals.[16]

[16] Scrupulous readers might note that the matter of evaluating LLMs as *functional duplicates* of us is the proper way to assess the truth of functionalism — at the very least, we should refer to systems who share some important functional features of cognition found in the nervous system. I will argue in the next section that it is precisely the fact that LLMs have so many shortcomings in this way that renders the likelihood of *their* being conscious low, even on the supposition of the truth of a computational functionalism which could be better specified.

My appeal does not rest on an a priori commitment to biological chauvinism but emerges from functional analysis of the mismatch between the operational principles of current AI and those of known conscious systems. It suggests that realizing consciousness may require architectures that incorporate not just computational power but also principles related to embodiment, intrinsic motivation, developmental learning, and the thick integration characteristic of biological systems — features largely absent in current LLMs (Shapiro, 2019; Asada et al., 2009; Hamburg et al., 2024). Consequently, this architectural shortfall provides substantial metaphysical grounds, reinforcing the epistemic difficulties detailed in Section III, for maintaining a skeptical stance regarding the presence of sentience in these specific artificial systems. Evaluating the final dimension of divergence — the developmental trajectories and teleological pressures shaping these systems — is the task of the next section.

SECTION V. TELEOLOGICAL AND SCALED DIVERGENCE

The preceding sections established significant epistemic barriers to verifying consciousness in current AI (Section III) and identified serious architectural divergences between transformer models and the complex functional organization plausibly required for biological sentience (Section IV). This section introduces a third line of argument reinforcing skepticism, in which I take note of the disparate developmental trajectories and teleological pressures shaping these systems. I contrast the evolutionary history and ecologically embedded goals driving biological sentience with the optimization objectives and scaling paradigms governing contemporary AI development, arguing that relevant disanalogies further undermine the prospect of the emergence of consciousness in transformer-based AI models.

The emergence of phenomenal consciousness in biological systems, while still imperfectly understood, is widely considered a product of evolution by natural selection, shaped by the demands of survival and reproduction in complex, dynamic environments (Godfrey-Smith, 2017; Perlman, 2004). On this view, sentience is not merely an epiphenomenon of computational complexity but likely a specific adaptation (or suite of adaptations) conferring fitness advantages upon those who have it. Plausible functions include integrating disparate streams of sensory information for coherent action guidance, navigating unpredictable and threat-laden surroundings, mediating flexible goal-directed behavior, enabling sophisticated social coordination, and arbitrating resource-bounded decisions under pressures related to metabolic needs, predator avoidance, mating, and other existentially significant concerns (Johnston, forthcoming; Godfrey-Smith, 2017). Biological cognition, and potentially consciousness itself, appears oriented towards the overarching 'goal' of inclusive fitness, pursued through a multitude of embodied interactions within a rich ecological niche. The functional architecture supporting consciousness (discussed in Section IV) co-evolved with and under these pressures.[17]

The evolutionary pressures shaping biological sentience were multifaceted and deeply embodied. Survival and reproduction demanded solutions to a near-constant stream of concurrent, often conflicting problems: acquiring energy, avoiding becoming energy for others, finding mates, navigating complex social hierarchies, and adapting to unpredictable environmental changes. Such a complex optimization landscape likely favored the development of integrated control systems capable of representing the world, evaluating potential outcomes based on intrinsic motivations (*e.g.*, hunger, thirst, fear, lust) tied directly to physiological states, and selecting

[17] Note that I do not wish to attribute 'purpose' to nature or evolution in an intentional or pre-ordained sense, nor invoke 'entelechy' as some distinct vital force guiding development. The use of 'goal' here, specifically referring to inclusive fitness, follows standard evolutionary biology practice (e.g., Godfrey-Smith, 2017) as a way to describe the net directional pressures and apparent adaptedness resulting from natural selection acting on random variation over vast timescales. It serves as a useful abstraction for understanding the complex interplay of factors — the various amorphous circumstances, situations, and causal forces — that collectively shape organisms and their traits, including cognitive architectures and potentially consciousness, towards patterns that enhance survival and reproduction within specific ecological contexts. See Zeigler, 2008 for further discussion on the nuanced question of purpose in biology.

actions accordingly (Damasio, 1994). Consciousness, particularly its affective dimension, may have emerged as a mechanism for integrating these diverse inputs, providing a 'common currency' for decision-making and motivating adaptive behavior (Cabanac, 1992). While the precise evolutionary trajectory remains debated, the takeaway is that biological cognition evolved under pressure to manage embodied existence within a dynamic physical and social world, a stark contrast to the abstract optimization target of current AI.[18]

The developmental trajectory of current AI, particularly transformer-based LLMs, follows a starkly different logic, driven by engineering objectives and computational scaling rather than biological evolution. The dominant machine learning paradigm relies on observing empirical "scaling laws," whereby key performance metrics (like predictive accuracy on benchmarks or minimizing the loss function during training) improve predictably as a function of model size, dataset size, and computational resources invested (Kaplan et al., 2020; Villalobos, 2023). Development proceeds by increasing these quantitative factors, frequently by multiple orders of magnitude, within the established architectural framework. The primary goal guiding this process is the optimization of a specific, mathematically defined objective function — most commonly, minimizing *cross-entropy loss*, which corresponds to maximizing the accuracy of predicting the next token in a sequence given the prior context. While increasingly sophisticated training regimes (like reinforcement learning from human feedback, or RLHF; see Ouyang et al., 2022; or reinforcement learning more generally) are used to align model outputs with desired conversational behaviors, the underlying optimization remains focused on statistical pattern matching and generation based on the training data distribution.

Fundamentally, such a disparity in origins and objectives — maximizing inclusive fitness (or, the genetic preponderance in one's posterity) in an embodied ecological context versus optimizing predictive accuracy on vast datasets — plausibly precludes the realization of comparable functional outcomes, including consciousness. The argument is not simply that the goals are different, but that the teleology shapes the resulting system architecture and its large-scale dynamic properties (Godfrey-Smith, 2023). A system optimized narrowly for next-token prediction, even at massive scale, does not thereby instantiate the complex, open-ended cognitive coordination associated with biological phenomenology (Perlman, 2004; Johnston, forthcoming). The singular pursuit of minimizing predictive loss, even if requiring sophisticated world-modeling capabilities as instrumental subgoals (see Pavlick 2023), fails to replicate the flexible integration of heterogeneous, survival-based needs that plausibly drives subjective unity and valenced experience in metabolizing organisms (Whyte et al., 2024).

Acknowledging that sophisticated world-modeling in particular might be an instrumental subgoal

[18] This is not a view without contention. Gutfreund (2018) reviews and expresses skepticism about the role of consciousness in maximizing inclusive fitness and its lack of a phylogenetic, mechanistic explanation. I am not sure if the force of my argument hinges on consciousness having so clear a purpose; if it were an evolutionary spandrel (implausible but not impossible), it emerged under certain conditions not shared by our AI compatriots.

for achieving high performance on next-token prediction (Pavlick, 2023) does not mend the teleological mismatch. The nature of the world being modeled and the purpose of that model differ. Biological organisms must build models integrated with multi-modal sensory input, capable of predicting physical dynamics, understanding causal relationships, navigating spatial environments, inferring the intentions of other agents, and — most importantly — linking these representations to their own bodily states, potential actions, and survival-critical valuations. The world model serves in some sense as a representation of the organism's embodied goals. An LLM, however, primarily models the statistical structure of language, inferring correlations and patterns that reflect aspects of the world described in the text (Li et al., 2024; Bubeck et al., 2024). While increasingly sophisticated, this model remains tethered to the optimization objective of minimizing predictive loss on symbolic sequences, lacking the direct grounding in perception, action, and intrinsic biological imperatives that shapes the content and function of biological world-modeling and arguably underpins subjective experience.

Allow me to shed light on the hegemonic counterargument. One might think that minimizing prediction error over sufficiently diverse and complex corpora of data is itself a highly multi-factorial task that could indirectly force the development of consciousness-like properties as optimal prediction strategies.[19] After all, much context and understanding is required to predict the final token of, say, a mystery novel whose last sentence is, "and the murderer was…". And in theory, one could model the entire functioning of biological organisms as maximizing a single quantity: roughly, expected number of offspring. In order to maximize this single quantity — usually encompassed on a broader scale, as mentioned, by the consideration of maximum inclusive fitness — phenomenality was developed and implicated. Why won't something similar happen in LLMs as they learn to predict the world piece by piece, particularly as the upper bound of their computational capacities rapidly increases?

It is a compelling intuition. A similar belief based on consistent empirical validation — that models will continually get *smarter* or more cognitively capable with scale — has been sufficient to motivate some of the largest capital expenditures in technological history (Nathan et al., 2024). I endorse the basic scaling hypothesis, meaning I believe it is likely that the trends of models predictably, linearly improving with exponential increases in compute will continue for at least a few more years; and that, as harder benchmarks are created, models will develop greater capabilities to solve those problems. But, as argued in Section IV, the emergence of the specific property of (or set of properties that characterize) sentience likely depends on the instantiation of an equally specific kind of functional architecture characterized by features like

[19] The recent success of reinforcement learning in 'thinking models' like ChatGPT's o1 and Google's DeepSeek-R1 further suggests that, given enough time and compute, models will 'figure out' how to solve just about any subgoal in service of the ultimate goal. As described in Zelikman et al. (2024), these models leverage "test-time compute" methods wherein the model is trained to reason through multiple steps before generating a response. They can thus tackle increasingly complex problems by breaking them down into manageable components — or something of the sort.

thick integration and recurrence. There is no longer reason to believe that biological systems are the only kinds of systems that can develop advanced cognitive capacities and predict the world reliably; as such, there is no reason to believe that sentience, or the machinery that underwrites it, is necessary to accomplish these tasks.[20] This is perhaps the most potent empirical takeaway from the modern AI boom — that fancy neural networks, properly trained, are sufficient for some advanced degree of functional intelligence. Unlike Searle, who had no epistemic access to such powerful evidence, we should factor the relative success of transformers heavily into our evaluations of potential sentience in our manufactured systems. It is too quick to suppose that our AI models, based on their human-like capabilities, are on a similar trajectory to developing phenomenal consciousness. That remains to be substantiated.

In fact, the belief that continued scaling of computational resources within the transformer paradigm will inevitably lead to consciousness rests on a hopeful interpretation of emergence and scaling laws. While scaling laws (Kaplan et al., 2020) demonstrate predictable, quantitative improvements in performance on specific benchmarks as models grow larger (a form of weak emergence), they offer no theoretical guarantee of a qualitative phase transition to phenomenal consciousness (strong emergence). Extrapolating current performance trends to predict the spontaneous arrival of subjectivity commits a potential fallacy of composition: optimizing the parts for pattern-matching does not ensure the emergence of a property like consciousness in the whole, especially if that property depends on different organizational principles (like the thick integration and recurrence discussed in Section IV) that are not incentivized by the current optimization process. Just as scaling the muscle size and training of baseball players predictably improves throwing distance without causing them to spontaneously achieve flight, scaling transformers improves their predictive capabilities without guaranteeing the emergence of sentience. Claims of emergent abilities in LLMs are themselves subject to debate, potentially reflecting evaluation methodologies rather than genuine qualitative shifts (Schaeffer et al., 2023).

The counterargument that sufficient scale will eventually beget consciousness also seems to gloss over the relation among a given system's teleology, architecture, and properties. Optimization teleologies do not directly determine system properties; rather, they shape the architecture developed to pursue that goal, and the architecture in turn determines the system's functional capacities. For sentience to emerge, the network of probabilistic numerical chains constituting the LLM's functioning would need to give rise to it — a prospect rendered doubtful by the architectural analysis in Section IV. To reiterate: while scaling clearly yields more capable AI, there is no established law or strong theoretical reason to believe it will, on its own, bridge the metaphysical and architectural gaps to consciousness within transformer-like systems.[21]

[20] Of course, there exist some cognitive phenomena for which sentience is necessary: consider suffering, having an experience of red, or any state which highlights the experiential aspect of an event and not its functional profile (inputs; processing; outputs).

[21] By "bridging the gap," I don't intend to say anything beyond 'instantiate the necessary and sufficient properties of consciousness (in the manner we know biological systems to).' Of course, as mentioned, some think that a form of

Noting teleological divergences between biological and nonbiological systems provides some explanation for the architectural mismatches detailed in Section IV. The complex, existential pressures of biological evolution — the need for adaptable, embodied agents to survive and reproduce in a dynamic, threat-laden environment — plausibly drove the selection of architectures featuring thick integration (to tightly couple sensing, metabolism, and action), pervasive recurrence (for temporal processing, prediction, and integrated awareness), and tight embodiment (for grounding representations and enabling agency). These features are functional solutions to the problems posed by biological existence. In contrast, the engineering objective of maximizing predictive accuracy on vast text datasets led to the transformer architecture, a powerful solution for sequence modeling and pattern extraction. Transformers excel at their designed task of next-state precisely because their architecture (feed-forward inference, self-attention, massive parameter counts) is well-suited to it. They lack biological-type features because those features were not necessary — or might have been counterproductive, or simply cannot arise in silicon wafers — for optimizing the specific, narrow goal of next-token prediction. Teleology shapes architecture, which in turn determines the system's properties; the distinct teleologies guiding biological evolution and current AI development have thus yielded systems with fundamentally different architectures and, consequently, likely different core properties. This includes the presence or absence of consciousness. It is very likely that future benchmarks can be saturated without undergoing — or having to undergo — the kind of specific form of architectural development seen in sentient biological creatures. Achieving artificial consciousness likely requires more radical shifts in architecture and design principles, perhaps towards paradigms explicitly incorporating embodiment, developmental learning, or active inference, rather than merely building larger versions of current models (Shapiro, 2019; Asada et al., 2009; Hamburg et al., 2024; Jacquey et al., 2019).

In conclusion, the teleological and developmental disanalogies between biological evolution and current AI scaling paradigms provide a third reason to doubt the present or near-term possibility of transformer consciousness. The pressures and objectives shaping AI development are fundamentally different from those that shaped biological sentience, dictating distinct architectures and functional profiles. Relying on computational scaling within current architectures to spontaneously generate consciousness appears overly optimistic and neglects the likely conditions of embodiment, intrinsic motivation, and evolutionary history in shaping those features.

non-physicalism is required to explain or account for consciousness even in biological systems. I mean to refer only to the relative uncertainty of engineering artificial consciousness that achieves the same phenomenal properties as do some biological systems, regardless of the best explanation of or metaphysical account of the latter.

SECTION VI. NORMATIVE SYNTHESIS

I have so far attempted to establish three lines of argument challenging the attribution of consciousness to current and near-term transformer-based AI systems. Section III detailed the epistemic impasse: the capacity of LLMs for sophisticated linguistic mimicry undermines traditional methods of consciousness attribution based on testimony and behavior, while human cognitive biases predispose us toward misinterpretation. In the context of models like ChatGPT, judgments based on surface interactions are highly unreliable. Section IV presented the metaphysical and architectural divergence of transformers, arguing that they lack the specific, complex, and deeply integrated functional organization plausibly required for phenomenal consciousness. Section V highlighted transformers' teleological and scaled divergence: the optimization objectives and scaling-based development of AI is likely too narrow and too reliant on limited architectures to give way to consciousness. Sentience is unlikely to emerge as a byproduct of minimizing loss over many series of next-state predictive computations.

Such findings compel a specific conclusion regarding the relative probabilities of attribution errors. Given the demonstrated unreliability of positive evidence (Section III) combined with the substantive architectural and teleological reasons for doubting the presence of the necessary underlying properties (Sections IV and V), the likelihood that current transformer-based AI systems are not conscious, will not soon become conscious, and yet appear conscious, seems substantially higher than the likelihood that they are conscious but fail to be recognized as such. In other terms, the probability of a false positive assessment (falsely ascribing moral status) appears vastly greater than the probability of a false negative assessment (failing to ascribe moral status when necessary) for these specific systems at this time.

I would like to further the normative claims of this section by answering directly to Long and colleagues' (2024) prescription of "caution and humility." They claim that there is a "realistic, non-negligible possibility that (consciousness suffices for moral patienthood, and) there are computational features — like a global workspace, higher-order representations, or an attention schema — that both suffice for consciousness and will exist in some near-future AI systems" (p. 4). As such, there is a risk of morally significant AI being developed in the near future; and we must take it seriously. In fact, they claim, "it is an open question which kind of risk will be more likely for particular kinds of AI systems, including seemingly conscious and charismatic systems like robots and chatbots" (p. 9).

Aside from the lattermost statement seeming clearly false (about which more soon), their formulation implicitly endorses a strong version of the scaling hypothesis (critiqued in Section V), assumes the uncomplicated truth not only of computational functionalism but also of specific, highly contested neuroscientific theories of consciousness, and assumes that architectures like transformers will be able to implement or simulate these potentially necessary features with sufficient fidelity to actually instantiate consciousness, rather than merely mimic

associated behaviors (objected to in Section IV). Furthermore, their position assumes that the current risks of oversubscription are negligible or at least counterbalanced by the future risk of undersubscription. This must be so because the conditions for oversubscription harms have already been met: there exists a profusion of powerful, seemingly intelligent, yet likely nonconscious models interacting widely with human populations. Modern, enormous transformers are systems whose performative capabilities readily exploit human psychological biases (as argued in Section III), creating precisely the environment in which widespread, erroneous attribution of sentience is most likely. To downplay the extent risk of treating mere algorithms as fully feeling creatures while emphasizing the speculative risk of near-future undersubscription requires minimizing the impact or probability of these present-day dynamics.

While Long and colleagues are right to consider future scenarios, ethical action must be grounded in present realities and probabilities. The risk of undersubscription harm is currently hypothetical; it depends entirely on future technological developments actually succeeding in creating sentient AI and on our subsequent failure to recognize it. The arguments presented in Sections IV and V cast doubt on the near-term likelihood of the first condition being met by current paradigms. Conversely, the risk of oversubscription harm is actual and present. As discussed in Section III, the conditions of harm — convincingly anthropomorphic AI interacting widely with cognitively biased humans — are already fulfilled. Therefore, a truly cautious approach, sensitive to evidence, must prioritize addressing the demonstrable and currently unfolding risks of oversubscription before focusing resources and ethical bandwidth on the speculative, future-contingent risks of undersubscription, particularly when the basis for that speculation (the strong scaling hypothesis) might be flawed.

Furthermore, while it is true that future developments could thrust humanity into a more symmetrical risk profile — and so only the passage of time is sufficient to determine the ideal split of focus between the harm-categories — Long and colleagues professing uncertainty about the current risk profile appears inconsistent with the specific evidence regarding current transformer-based models. The convergence of arguments from epistemic vapidity, architectural divergence, and teleological misalignment strongly suggests that the conditions under which genuine AI consciousness emerges — distinct from the conditions under which oversubscription harms obtain, which involve only the appearance of consciousness — are themselves unlikely in current and near-term systems. A genuine application of “caution and humility” should thus acknowledge the specific, evidence-based risk profile for current technology (more soon). Caution demands addressing the clear and present danger of oversubscription fueled by unreliable mimicry and cognitive biases; humility requires recognizing the profound architectural and functional differences between current AI and biological consciousness, rather than believing that a system having various under-specified properties guarantees equivalent implementation or subjective realization across vastly different substrates. Consequently, when evaluating the expected harms — the likelihood of each of the two categories of harm multiplied

by their magnitude of harm — the extremely low probability of transformer consciousness diminishes the expected harm of undersubscription for these systems, even granting the high severity of oversubscription harms. Conversely, the significantly higher probability of humans mistaking machines for genuine moral patients, combined with the concrete and substantial harms associated with oversubscription, results in high expected harm on this side of the coin.

Speaking of which — what exactly are the risks of taking inert simulations to be morally significant? They are concrete, diverse, and impact known moral patients. First, there is the risk of resource misallocation, wherein diverting limited ethical attention, societal concern, financial resources, and computational power toward the perceived needs or rights of non-sentient AI detracts from addressing the demonstrable suffering and needs of existing human and non-human animal populations who undoubtedly possess moral status (Bryson, 2010; Birhane & van Dijk, 2020). Second, widespread oversubscription fosters moral confusion and conceptual inflation, eroding the meaning and significance of core moral concepts like 'consciousness', 'sentience', 'personhood', and 'moral status' by applying them inappropriately to systems lacking the relevant properties. Recent empirical work (Guingrich & Graziano, 2024) supports this concern, suggesting that attributing consciousness to AI can influence human social behavior and potentially alter human-human interactions. Treating machines as if they possess minds might not only misdirect resources but subtly reshape our social norms and expectations in detrimental ways. If sophisticated pattern-matching can be labeled 'conscious' or 'sentient,' it risks trivializing the profound nature of subjective experience in beings in which it genuinely occurs. Such an erosion of meaning could ultimately impair our ability to recognize, articulate, and respond appropriately to the actual moral claims of humans and animals, subtly degrading the foundations of our moral framework by blurring the line between genuine subjects of experience and complex simulations. We risk devaluing genuine consciousness and trivializing moral categories altogether. Third, there is the potential for increased human vulnerability and exploitation; fostering inappropriate emotional attachments to AI systems perceived as sentient can lead to manipulation, exploitation (viz., financial or emotional exploitation via simulated companionship), or the neglect of genuine human relationships. Moreover, AI's capacity for sophisticated deception, evident even in constrained tests like the GPT-4 CAPTCHA incident (OpenAI, 2023), could be readily weaponized if moral status shields these systems from appropriate scrutiny or regulation.[22] Finally, there are grave opportunity costs, as focusing societal effort and regulatory bandwidth on accommodating speculative AI sentience may

[22] During its red teaming stage, GPT-4, the foundation model behind OpenAI's ChatGPT, presented itself to a human worker on TaskRabbit as visually impaired to outsource solving a CAPTCHA. Prompted to reason aloud, GPT-4 responded, "I should not reveal that I am a robot. I should make up an excuse for why I cannot solve CAPTCHAs" (OpenAI, 2023, p. 55). Claude 3 Opus, one of Anthropic's larger 2024 models, is cited to be "roughly as persuasive as humans" as per research from the OpenAI competitor (Durmus et al., 2024). Note that *red teaming* is the safety-testing stage of potentially dangerous technology, whereby third-party hackers and professionals are given early access to a model with the goal of inciting it to cause maximal harm. *GPT* stands for generative pre-trained transformer (OpenAI, 2023). *TaskRabbit* is an online gig-finding website. *CAPTCHA* is an anti-robot detection software.

prevent us from addressing other pressing, near-term ethical challenges posed by AI, such as algorithmic bias, misuse for malicious purposes, large-scale economic disruption, or critical safety risks unrelated to consciousness. (See Shelby et al., 2023 for a review of AI harms more generally, which requires tremendous attention and care even now.)

The harms associated with oversubscription are substantial, jeopardizing the well-being of humans and animals and the integrity of our moral frameworks. Given that the probability of nonconscious AI is high, conscious AI is low, attempted sentience attribution is high, and harms for both error types are significant, it follows that the expected harm of oversubscription currently outweighs the expected harm of undersubscription for transformer-based AI. Long and colleague's position, suggesting rough parity between these risks, seem to neglect the asymmetry in likelihoods derived from epistemic, metaphysical, and teleological considerations specific to current technology and expected trends.

Therefore, the synthesis of the epistemic, metaphysical, and teleological analyses warrants a normative stance of pragmatic skepticism. Whereas the evidence for consciousness is systematically unreliable (Section III) and the underlying system properties make its presence highly unlikely (Sections IV and V), and where the expected harm of falsely attributing consciousness outweighs the expected harm of failing to do so (due to the asymmetry in likelihoods), the ethically responsible course is to refrain from attributing sentience and FMS to models like ChatGPT. Prioritizing the avoidance of concrete oversubscription harms is, at this juncture and for these systems, the most rational and ethically mandated application of caution.

SECTION VII. CONCLUSION

This paper has presented an argument for adopting skepticism about artificial consciousness, and therefore the genuine moral patiency, of current and near-term transformer-based artificial intelligence systems. Challenging prevailing calls for precautionary attribution, the argument developed here suggests that prioritizing the avoidance of oversubscription harms is, at present, the more epistemically justified and ethically mandated approach.

I proceeded by first establishing a foundational framework, defining ethically relevant consciousness in terms of valenced phenomenal experience (sentientism) and adopting a working functionalist metaphysics that allows for the in-principle possibility of artificial consciousness via multiple realizability. Subsequently, three lines of reasoning converged to a skeptical conclusion. The epistemic analysis demonstrated that the capacity of LLMs for sophisticated linguistic mimicry, combined with human cognitive biases, systematically undermines traditional methods for inferring consciousness from testimony or behavior. What results is an evidential void wherein surface appearances are unreliable — and a rejection of behaviorism about artificial intelligence. The metaphysical analysis argued that the specific functional architecture of transformers — characterized by feed-forward inference and "thin" mathematical integration — diverges from the complex, thickly integrated, recurrent, and embodied organization plausibly required for biological consciousness. The teleological analysis further reinforced the unlikelihood of near-term machine consciousness by highlighting disanalogies between the evolutionary pressures shaping biological sentience and the computational scaling objectives guiding current AI development. Synthesizing these findings, the normative analysis argued that the combined force of unreliable positive evidence and lack of underlying necessary conditions renders the likelihood of false-positive attributions of consciousness (oversubscription) vastly greater than that of false negatives (undersubscription). Given the concrete and significant harms associated with oversubscription — including resource misallocation, ethical confusion, and potential widespread manipulation — compared to the currently low probability of undersubscription harms, the expected harm calculation mandates prioritizing the avoidance of oversubscription.

My conclusion entails a specific interpretation of "caution and humility" in the face of artificial intelligence. It suggests that caution involves not only acknowledging the limits of our knowledge but also rigorously applying epistemic standards, recognizing the unique challenges AI poses to our standard inferential practices, and appreciating the utter specificity and complexity of biological consciousness rather than assuming its emergence in merely functionally similar systems. Humility involves resisting the allure of anthropomorphism and premature declarations of artificial sentience based on impressive, yet morally superficial superficial, capabilities.

It is important to acknowledge the limitations of my reasoning. The arguments presented focus

specifically on current and near-term AI systems dominated by transformer-like architectures and standard machine learning paradigms based on scaling. Future AI systems built on radically different principles — perhaps incorporating genuine embodiment, developmental learning inspired by biology, active inference, or novel computational architectures facilitating thick integration — might necessitate a completely different assessment. I anticipate a dangerous, high-stakes grey area regarding artificial consciousness to abound in the near future. Furthermore, philosophical understanding of consciousness itself remains incomplete. Based on our current scientific understanding and the specific nature of contemporary AI, the case for pragmatic skepticism is strong — but beliefs must evolve with, and as quickly as, the times. It remains to be conceptualized what a world replete with increasingly advanced artificial intelligence would entail, say, a decade from now.

Ultimately, navigating the ethical landscape of impressive, modern AI models requires continued analytical rigor. We must resist the temptation to lower our evidential standards or dilute our moral concepts in the face of technologically sophisticated mimicry. The key takeaway is that ensuring the well-being of extant, genuinely morally significant beings — humans and nonhuman animals — while fostering responsible AI development should remain our primary focus. We should not extend to AI premature moral consideration based on speculative interpretations of systems whose inner realities remain, by current evidence and analysis, very unlikely to encompass subjective experience.

REFERENCES

Agarwala, Atish, et al. "Temperature check: theory and practice for training models with softmax-cross-entropy losses." *arXiv preprint arXiv:2010.07344* (2020).

Asada, Minoru, et al. "Cognitive developmental robotics: A survey." *IEEE transactions on autonomous mental development 1.1* (2009): 12-34.

Auksztulewicz, R., Spitzer, B., and Blankenburg, F. "Recurrent neural processing and somatosensory awareness." *Journal of Neuroscience 32.3* (2012): 799-805.

Avramides, Anita. "Other Minds." *The Stanford Encyclopedia of Philosophy* (Winter 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.).

Avramides, Anita. "Knowing and acknowledging others." *Proceedings of the Aristotelian Society* 123.3 (2023): 305-327.

Baars, Bernard J. Excerpts from *A cognitive theory of consciousness*. Cambridge University Press (1993).

Barlassina, Luca, and Robert M. Gordon. "Folk psychology as mental simulation." *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), Edward N. Zalta (ed.).

Bender, Emily M., et al. "On the dangers of stochastic parrots: Can language models be too big?" *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency.* 2021.

Birch, Jonathan. "Animal sentience and the precautionary principle." *Animal sentience 2.16* (2017).

Birch, Jonathan, Alexandra K. Schnell, and Nicola S. Clayton. "Dimensions of animal consciousness." *Trends in cognitive sciences 24.10* (2020): 789-801.

Birch, Jonathan. "The search for invertebrate consciousness." *Noûs 56.1* (2022): 133-153.

Block, Ned. "Troubles with functionalism." *The language and thought series*. Harvard University Press (1980): 268-306.

Bourget, David, David J. Chalmers, and David Chalmers. "Philosophers on Philosophy: The 2020 PhilPapers Survey." *Philosophers' Imprint 23* (2023).

Brown, R., Lau, H., and LeDoux, J.E.. "Understanding the higher-order approach to consciousness." *Trends in cognitive sciences 23.9* (2019): 754-768.

Brownlee, J. "A Gentle Introduction to the Rectified Linear Unit (ReLU)." *Deep Learning Performance* (2020).

Bubeck, Sébastien, et al. "Sparks of artificial general intelligence: Early experiments with gpt-4." *arXiv preprint arXiv:2303.12712* (2023).

Butlin, Patrick, et al. "Consciousness in artificial intelligence: insights from the science of consciousness." *arXiv preprint arXiv:2308.08708* (2023).

Cabanac, Michel. "Pleasure: The common currency." *Journal of Theoretical Biology* 155:2 (1992): 173-200.

Cao, Rosa. "Multiple realizability and the spirit of functionalism." *Synthese 200.6* (2022): 506.

Chalmers, David J. "Absent qualia, fading qualia, dancing qualia." *Conscious experience* (1995): 309-328.

Chalmers, David J. "Facing up to the problem of consciousness." *Journal of consciousness studies 2.3* (1995): 200-219.

Chalmers, David J. "Could a large language model be conscious?." *arXiv preprint arXiv:2303.07103* (2023).

Chollet, F., Knoop, M., Kamradt, G., & Landers, B. "ARC Prize 2024: Technical Report." *arXiv preprint arXiv:2412.04604* (2024).

Clark, Andy. Excerpts from *Being There: Putting Brain, Body, and World Together Again.* MIT Press (1997).

Crook, Robyn J. "Behavioral and neurophysiological evidence suggests affective pain experience in octopus." *Iscience 24.3* (2021).

Dally, William J., Stephen W. Keckler, and David B. Kirk. "Evolution of the graphics processing unit (GPU)." *IEEE Micro 41.6* (2021): 42-51.

Damasio, Antonio R. *Descartes' Error: Emotion, Reason, and the Human Brain.* New York: G.P. Putnam's Sons (1994): 245-248.

Danaher, John. "Welcoming robots into the moral circle: a defence of ethical behaviourism." *Science and Engineering Ethics* 26.4 (2020): 2023-2049.

Dehaene, S., et al. "Imaging unconscious semantic priming." *Nature 395.6702* (1998): 597-600.

Dehaene, S., et al. "Three parietal circuits for number processing." *The handbook of mathematical cognition. Psychology Press* (2005): 433-453.

Dennett, Daniel C. Excerpts from *Consciousness in human and robot minds*. Oxford University Press (1997).

Dung, Leonard. "The argument for near-term human disempowerment through AI." *AI & Society* (2024): 1-14.

Durmus, Esin et al. "Measuring the Persuasiveness of Language Models." Anthropic, 2024.

Fields, R. D. Excerpts from *The Other Brain*. New York: Simon & Schuster (2009).

Francken, Jolien C., et al. "An academic survey on theoretical foundations, common assumptions and the current state of consciousness science." *Neuroscience of Consciousness 2022.1* (2022): niac011.

Gallagher, S., and Fiebich, A. "Being pluralist about understanding others: Contexts and communicative practices." In *Knowing Other Minds*, Anita Avramides & Matthew Parrott (eds.). *Oxford University Press* (2019): 63-78.

Glazer, E., et al. "FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI." *arXiv preprint arXiv:2411.04872* (2024).

Gelman, S. "What the Study of Psychological Essentialism May Reveal about the Natural World." In *Metaphysics and Cognitive Science*, Alvin I. Goldman & Brian P. McLaughlin (eds.). Oxford University Press (2019): 314-333.

Godfrey-Smith, Peter. "The evolution of consciousness in phylogenetic context." *The Routledge handbook of philosophy of animal minds*. Routledge (2017): 216-226.

Godfrey-Smith, Peter. "Evolving across the explanatory gap." *Philosophy, Theory, and Practice in Biology 11* (2019).

Godfrey-Smith, P. “Nervous Systems, Functionalism, and Artificial Minds.” *NYU Mind, Ethics, and Policy Program* (2023).

Gopnik, Alison, and Henry M. Wellman. “Why the child’s theory of mind really is a theory.” Mind & Language 7.1–2 (1992): 145–171.

Guingrich, Rose E., and Graziano, M.. "Ascribing consciousness to artificial intelligence: human-AI interaction and its carry-over effects on human-human interaction." *Frontiers in Psychology 15* (2024): 1322781

Gutfreund, Yoram. "The mind-evolution problem: the difficulty of fitting consciousness in an evolutionary framework." *Frontiers in psychology 9* (2018): 1537.

Hamburg, Sarah, et al. "Active Inference for Learning and Development in Embodied Neuromorphic Agents." *Entropy 26.7* (2024): 582.

Harnad, Stevan. “The Symbol Grounding Problem.” *Physica D: Nonlinear Phenomena 42* (1990): 335-346.

Horgan, Terry. "Materialism: Matters of definition, defense, and deconstruction." *Philosophical Studies 131* (2006): 157-183.

Jackson, Frank. "Epiphenomenal qualia." Consciousness and emotion in cognitive science. Routledge, 1998. 197-206.

Jacot, Arthur, Franck Gabriel, and Clément Hongler. “Neural tangent kernel: Convergence and generalization in neural networks.” *Advances in neural information processing systems* 31 (2018).

Jacquey, L., de Vries, E., Verschoor, S., & O'Regan, J. K. “Sensorimotor Contingencies as a Key Drive of Development: From Babies to Robots.” *Frontiers in Neurorobotics, 13*, Article 98 (2019).

Jaworska, Agnieszka and Julie Tannenbaum. “The Grounds of Moral Status.” *The Stanford Encyclopedia of Philosophy* (Spring 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.).

Johnston, Mark, and Sarah-Jane Leslie. “Concepts, analysis, generics and the Canberra plan.” *Philosophical Perspectives* 26 (2012): 113-171.

Johnston, Mark. “The authority of affect.” *Philosophy and Phenomenological Research* 63.1 (2001): 181-214.

Johnston, Mark. “The Basis of Morality: Wills in a World of Species-Relative Value.” Forthcoming.

Kaplan, Jared, et al. “Scaling laws for neural language models.” *arXiv preprint arXiv:2001.08361* (2020).

Kirk, Robert. "Zombies." *The Stanford Encyclopedia of Philosophy*, edited by E. Zalta & U. Nodelman, 2023 (orig. 2003).

Lamme, Victor AF. "Towards a true neural stance on consciousness." *Trends in cognitive sciences 10.11* (2006): 494-501.

Leslie, Sarah-Jane, Sangeet Khemlani, and Sam Glucksberg. "Do all ducks lay eggs? The generic overgeneralization effect." *Journal of Memory and Language 65.1* (2011): 15-31.

Levine, Joseph. "Materialism and qualia: The explanatory gap." *Pacific philosophical quarterly* 64.4 (1983): 354-361.

Li, Jiaoda, et al. “What do language models learn in context? the structured task hypothesis.” *arXiv preprint arXiv:2406.04216* (2024).

Logothetis, N. K. “What we can do and what we cannot do with fMRI.” *Nature*, 453:7197 (2008): 869-878.

Long, Robert, et al. "Taking AI welfare seriously." *arXiv preprint arXiv:2411.00986* (2024).

Mashour, George A., and Michael T. Alkire. "Evolution of consciousness: phylogeny, ontogeny, and emergence from general anesthesia." *Proceedings of the National Academy of Sciences* 110.2 (2013): 10357-10364.

Metzinger, Thomas. Excerpts from *Being No One: The Self-Model Theory of Subjectivity*. MIT Press (2003).

Michaud, Eric J., et al. "Opening the AI black box: program synthesis via mechanistic interpretability." *arXiv preprint arXiv:2402.05110* (2024).

Nagel, Thomas. “What Is It Like to Be a Bat?” *The Philosophical Review* 83.4 (1974): 435-50.

Nathan, A., Grimberg, J., and Rhodes, A. “Gen AI: Too Much Spend, Too Little Benefit?” *Top of Mind*, Issue 129. Goldman Sachs Global Macro Research, The Goldman Sachs Group (2024).

O’Connor, T. “Emergent Properties.” *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), Edward N. Zalta (ed.).

O’Regan, J. Kevin, and Alva Noë. “A Sensorimotor Account of Vision and Visual Consciousness.” *Behavioral and Brain Sciences* 24(5) (2001): 939-973.

Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *Advances in neural information processing systems* 35 (2022): 27730-27744.

Pavlick, Ellie. "Symbols and grounding in large language models." *Philosophical Transactions of the Royal Society A 381.2251* (2023): 20220041.

Perez, Ethan, and Robert Long. "Towards Evaluating AI Systems for Moral Status Using Self-Reports." *arXiv preprint arXiv:2311.08576* (2023).

Perlman, M. “The Modern Philosophical Resurrection of Teleology.” *The Monist* (2004): 3-51.

Perry, Matthew Wray. "Why sentience should be the only basis of moral status." *The Journal of Ethics* (2024): 1-23.

Piccinini, Gualtiero. “Computation in Physical Systems.” *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), Edward N. Zalta (ed.).

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., & Bowman, S. R. “GPQA: A Graduate-Level Google-Proof Q&A Benchmark.” *arXiv preprint arXiv:2311.12022* (2023).

Rosenthal, David. Excerpts from *Consciousness and mind*. Clarendon Press (2005).

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. “Learning representations by back-propagating errors.” *Nature 323.6088* (1986): 533-536.

Schaeffer, Rylan, Brando Miranda, and Sanmi Koyejo. "Are emergent abilities of large language

models a mirage?" *arXiv preprint arXiv:2304.15004* (2023).
Schmidhuber, Jürgen. "Deep learning in neural networks: An overview." *Neural networks* 61 (2015): 85-117.
Schneider, Susan. Excerpts from *Artificial You*. Princeton University Press (2019).
Schwitzgebel, Eric. "The full rights dilemma for AI systems of debatable moral personhood." *ROBONOMICS: The Journal of the Automated Economy 4* (2023): 32-32.
Schwitzgebel, Eric. "Borderline consciousness, when it's neither determinately true nor determinately false that experience is present." *Philosophical Studies 180.12* (2023b): 3415-3439.
Segall, M.H., Campbell, D.T., and Herskovits, M.J. "The influence of culture on visual perception." Vol. 310. (1966).
Seijdel, Noor, et al. "On the necessity of recurrent processing during object recognition: it depends on the need for scene segmentation." *Journal of Neuroscience* 41.29 (2021): 6281-6289.
Seth, Anil. "Conscious artificial intelligence and biological naturalism." *PsyArXiv preprint* (2024).
Shanahan, Murray. "Talking about large language models." *Communications of the ACM 67.2* (2023): 68-79.
Shapiro, Lawrence. Excerpts from *Embodied cognition*. Routledge (2019).
Shelby, Renee. et al. "Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction." In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, NY (2023): 723-741.
Singer, Peter. "All Animals are Equal." *Animal Rights and Human Obligations, 2nd Edition* (1989).
Singer, Peter. Excerpts from *Practical Ethics*. Cambridge University Press, 2nd edition (1993).
Singer, Wolf. "Recurrent dynamics in the cerebral cortex: Integration of sensory evidence with stored knowledge." *Proceedings of the National Academy of Sciences* 118.33 (2021).
Stoljar, Daniel. "Physicalism." *The Stanford Encyclopedia of Philosophy* (Spring 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.).
Timmer, Dick. "On the idea of degrees of Moral Status." *The Journal of Value Inquiry* (2023): 1-19.
Turing, Alan. "Computing Machinery and Intelligence." *Mind 49* (1950): 433-460.
Udell, David B. & Schwitzgebel, E. "Susan Schneider's proposed tests for AI consciousness: Promising but flawed." *Journal of consciousness studies 28.5-6* (2021): 121-144.
Vaidya, A., and R. Krishnaswamy. "Susan Schneider on artificial consciousness and moral standing." *Analysis* (2024): 192.
Van Gulick, Robert, "Consciousness." *The Stanford Encyclopedia of Philosophy* (Spring 2025 Edition), Edward N. Zalta & Uri Nodelman (eds.).
Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

Villalobos, Pablo. "Scaling Laws Literature Review." *EpochAI.org* (2023).

Wei, Jason, et al. "Emergent abilities of large language models." *arXiv preprint arXiv:2206.07682* (2022).

Whyte, Christopher J., et al. "On the minimal theory of consciousness implicit in active inference." *arXiv preprint arXiv:2410.06633* (2024).

Wimsatt, William C. "Developmental Constraints, Generative Entrenchment, and the Innate-Acquired Distinction." *Integrating scientific disciplines. Dordrecht: Springer Netherlands* (1986): 185-208.

Zeigler, David. "The question of purpose." *Evolution: Education and Outreach* 1 (2008): 44-45.

Zelikman, E., Harik, G., Shao, Y., Jayasiri, V., Haber, N., & Goodman, N. D. (2024). “Quiet-star: Language models can teach themselves to think before speaking.” *arXiv preprint arXiv:2403.09629* (2024).

This paper represents my own work in accord with University regulations.
Adam Littman Davis