Toward Criteria for Artificial Self-Consciousness: Unity, Normativity, and Agency

B. Scot Rousse
UC Berkeley
bsrousse@gmail.com

Abstract

This paper distinguishes two forms of consciousness that are relevant for debates about artificial intelligence: pre-reflective experiential awareness and reflective self-consciousness. Drawing on phenomenology, Kant, and contemporary philosophy of mind, it argues that pre-reflective awareness involves the minimal self-involvement characteristic of phenomenal experience, while reflective self-consciousness involves a unified standpoint from which a subject can take a stand about how things are, evaluate them under norms of truth and value, and revise them in light of reasons. The paper analyzes reflective self-consciousness in terms of agency, normativity, and unity, articulating a structure of epistemic answerability that includes commitment formation, persistence across time, conflict detection, and revision in response to error. Distinguishing these two forms of self-involvement illuminates the unstable ethical landscape of artificial consciousness and suggests that emerging artificial systems may pressure the inherited moral categories through which moral standing has traditionally been understood.

Self-awareness and Self-Consciousness

Asking which theories of consciousness apply to artificial systems requires asking what kinds of awareness the theories are supposed to be modeling. Contemporary discussions of consciousness in AI frequently elide an important distinction: that between pre-reflective self-awareness, often called experiential or phenomenal consciousness, and reflective self-consciousness. The latter pertains to an experiencing subject's agentic capacity *to make up its own mind* regarding what to think and do; to assume a unified normative standpoint from which one forms, evaluates, and revises one's own beliefs and intentions. This is a distinct mode of being-involved in one's own experiences. On the view presented here, human cognition is not an isolated internal process. It is constitutively shaped by situated engagement with the world, such that belief formation, revision, and action remain responsive to environmental constraint as well as to internal representational structure.

Such a finer-grained picture of human consciousness helps clarify the moral stakes of attempts to engineer artificial consciousness; it also intensifies the question of whether our inherited moral categories are adequate to the novel kinds of minds that may one day emerge (Schwitzgebel 2023).

Much contemporary research on consciousness in AI is focused on phenomenal consciousness, the "something it is like" character of an experience (Nagel 1974, Block 1995). For example, a notable recent survey (Butlin et al 2023) explicitly focuses only on phenomenal consciousness. The authors argue that phenomenal consciousness (when combined with the contested assumption of computational functionalism) is tractable by a range of contemporary neuroscientific theories that also have application to artificial intelligence research (e.g., recurrent processing, global workspace, and higher-order theories).

Notwithstanding the prima facie plausibility of this methodological fastidiousness, research into AI consciousness will be compromised unless it goes beyond a narrow focus on phenomenal consciousness. Any comprehensive attempt to map the space of conscious awareness must also attend to reflective self-consciousness. This experiential capacity lies at the center of our moral agency and practices of mutual accountability. Recognizing this distinction is therefore essential for thinking clearly about the moral dilemmas connected with the possibility of artificial consciousness.

Schwitzgebel and Garza (2015, 2020), for example, argue that developers face a "design dilemma": either construct systems that clearly fall short of moral status, or deliberately design systems capable of personhood and treat them accordingly from the outset. Other approaches emphasize precaution. Sebo (2020) and others in the discussion about AI welfare (e.g., Metzinger 2021; Long et al. 2024), argue that if there is a non-negligible probability that artificial systems could possess morally relevant forms of experiential con-

sciousness, developers should act under conditions of ethical precaution. The distinction developed in this paper complements these proposals by clarifying the form of self-involvement that might underlie these different forms of moral standing. Distinguishing pre-reflective experiential consciousness from reflective self-consciousness helps locate more precisely where precautionary obligations may arise and where the stronger demands associated with moral agency would begin to apply. However, in the end, pressing the question of how far the distinctions introduced here apply to new forms of agency that may emerge deepens the impression that advances in AI will fundamentally destabilize our inherited moral intuitions and categories.

Recent scholars of phenomenology (Rousse 2019; Zahavi 2005; Zahavi 2014) have shown that the figures in this tradition (e.g., Husserl, Heidegger, Sartre, Merleau-Ponty) made important contributions to the theorizing of what Block (1995) later called "phenomenal consciousness." I follow Zahavi (2005) in using the labels "pre-reflective self-awareness" or "experiential consciousness" as these terms help to set up the contrast to "reflective self-consciousness." Pre-reflective self-awareness does not involve judgment or deliberation; it is not a matter of observing or evaluating one's experience. It includes at least five interrelated characteristics. Pre-reflective experiential consciousness is a form of minimal self-involvement that:

1. Amounts to the basic being awake or "there" in one's experience such that there is something the experience is like.
2. Precedes, grounds, and can be taken up in, episodes of reflection, as when we focus on some particular content or facet of our perceptual experience.
3. Is "non-objectifying" or what Sartre (2018) calls "non-positional." That is, pre-reflective self-awareness does not take or "posit" itself as a focal object of awareness; it does not turn back upon itself as another thing we are aware of. It is a form of implicit or liminal self-awareness correlative to the consciousness of anything else.
4. Asymmetrically presents and unifies experience as *my* experience, not your experience.
5. Transpires as a stream or temporal flow

If artificial systems were ever somehow to instantiate such a minimal form of experiential consciousness, the ethical stakes would intensify and the kinds of precautionary considerations discussed in recent work on AI welfare would begin to apply (Birch 2025; Metzinger 2021; Long et al. 2024; Sebo 2025). The plausibility of such a scenario currently depends on the dubious assumption of computational functionalism about consciousness (Seth 2026; Block 2025). Ultimately, the truth of computational functionalism is an empirical matter. But given the uncertainty involved, the pace of technological change and the proliferation of ethical dilemmas, the relevant conceptual and moral clarifications cannot await scientific resolution of the facts.

Metzinger (2020, 2021) speculates that characteristics 1, 2, and 3 above could obtain jointly and separately from characteristic 4. He calls this "minimal phenomenal experience" and contends that it would instantiate a non-egoic form of awareness. Metzinger argues further that characteristic 4, the fact that experience is given as *my* experience, is necessary for the experience of suffering and thus, on some views, for being considered a plausible moral patient. We will return to such issues in the final section.

In any case, an all-important feature of experiential consciousness is its unified character (Bayne and Chalmers 2003). Leading neuroscientific and philosophical theories of consciousness converge on the idea that an essential feature of conscious experience is its felt and functional unity or integration, even if they disagree about the mechanisms that produce this.

One dimension is synchronic unity: the integration of diverse informational, perceptual, cognitive, and affective processes into a single experiential field at a given moment. As Butlin et al. (2023) report, global Workspace theories explain this through the broadcasting of information across otherwise specialized neural subsystems (Baars 1988; Dehaene et al. 1998; Mashour et al., 2020), while recurrent processing accounts emphasize feedback loops that bind perceptual features into unified and coherent visual scenes (Lamme 2010, 2020). Midbrain theory claims that activity in the midbrain and basal ganglia enables spatiotemporal modeling that integrates spatial, affective, and homeostatic information into a "unified, multimodal neural model of the agent within its environment" (Klein and Barron 2016).

A second dimension is diachronic unity: the organization of experience across time into a continuous stream in which earlier and later moments belong to the same unfolding unit of awareness. Recurrent processing, memory, and self-modeling architectures have all been proposed as mechanisms that stabilize and extend conscious contents across successive moments (Butlin et al 2023; Phillips 2018; Klein and Barron 2016).

Today's LLM-based AI systems are highly unlikely to support either of these forms of experiential unity (Butlin et al 2023; Birch 2025; Chalmers 2023). Yet, even if we grant the possibility that future AI systems might in principle instantiate mechanisms sufficient for minimal pre-reflective self-awareness, this would not endow them with the capacities involved in the reflective self-consciousness characteristic of and available to human subjects.

Reflective Self-Consciousness: "I Think"

According to a tradition running from Kant and Frege through classical phenomenology and contemporary analytic philosophy of mind, one special experiential capacity of the human subject is our ability actively to make up and change our own minds regarding what to believe and do. These are the marquee experiences of reflective self-consciousness, enabling us to take up a first-person, participant's perspective on our own mental and motivational life, rather than just being a specially placed theoretical observer with "privileged access" to a stream of experience or contents of a mind that happens to be mine.

I will now briefly sketch three interrelated dimensions of reflective self-consciousness: agency, normativity, and unity, beginning with their role in the question of what to think (theoretical reason), before briefly touching on their role in the question of what to do (practical reason).

Agency: Making Up Your Mind

A central feature of reflective self-consciousness is the capacity to take a practical, agential involvement in the make-up of one's own mind. Not every system capable of experiences has a participant's, first-person relation to its own mental life. A subject capable of reflective self-consciousness, however, is more than a passive site where experiences play out, beliefs arise, or integrated information flows.

The capacity for reflective self-consciousness involves the capacity for *making up* or *changing* one's own mind about what to think and do; that is, for exercising practical, first-person authority over one's propositional attitudes, such as beliefs and desires. When a person answers the question whether they believe that *p*, they usually do not undertake a theoretical inquiry directed at a psychological or computational state to be discovered through inner observation (Moran 2001, Boyle 2024). Instead, they determine whether they believe that *p* by looking out at the world and seeing whether *p* is true. This authority also extends to the active revision of beliefs in light of a situation evolves.

This phenomenon is often described in terms of the transparency of belief. The question "Do I believe that *p*?" is ordinarily answered by looking out to the world and seeing whether *p*. The self-directed question about one's belief is thus transparent to the world-directed question about the relevant state of affairs (Moran 2001; Boyle 2024). In reflecting on the latter, the subject exercises its rational capacity for judgment, the capacity to form or revise a commitment to the truth of a proposition. My beliefs are therefore not normally fixed psychological facts awaiting discovery.

Of course, not all of our attitudes stand under this kind of authority. Some beliefs occur in ways that we experience as alien, given facts, impervious to our reflection upon how things are. A person might discover, for example, that they harbor an implicit bias only by observing their own behavior, even though upon reflection they reject the belief revealed by that behavior (Boyle 2024). In such cases the estranged belief is attributed to oneself as a psychological fact rather than endorsed through deliberation. But overall, to be capable of reflective self-consciousness is to be able to occupy an agentic, first-person perspective on one's beliefs, to own them rather than merely find or host them, to revise them in light of how things are in the world, and to answer for them when called upon. This goes beyond higher-order observation of one's beliefs: it is a matter of a self-constituting power of the mind, the power to *make up* and change our minds rather than just being an expert witness to them (Moran 2001).

A well-known cognitive scientist once pushed back on me after hearing an argument like the one sketched here. "There are two kinds of philosophers of mind," he said, "those who need to take more drugs and those who need to take less." The line is at once a funny joke and a thought-provoking philosophical jab, but it missed my point, even while inadvertently reinforcing it. If certain altered states loosen the subject's grip on its own attitudes, that does not show that reflective self-consciousness is irrelevant. It shows what drops out when that capacity is compromised. The stream of thought and experience continues, but the subject's ability to stand behind its attitudes, revise them, and answer for them begins to come apart. If AI research eventuates in artificial minds that will participate in our worlds, our institutions, and our practices of care, then machine consciousness research cannot postpone dealing with the capacities involved in reflective self-consciousness. Those capacities are integral to our practices of shared scientific inquiry and mutual accountability.

Normativity: Aiming at Truth

When a reflectively self-conscious subject determines what it believes, it aims at truth. Belief is not merely a psychological or representational state that gets triggered in a system. It is a commitment to a proposition concerning how things are in the world, a commitment that can succeed or fail, and that the subject is motivated to revise or revoke, depending on whether the world is as the belief represents it to be.

Frege's distinction between thoughts and mere psychological occurrences drives this point home (Frege 1997). Experiences such as sense impressions, images, feelings, or associative trains of thought may occur within the stream of experience without committing the subject to anything, or without the subject actively judging that anything is *so*. These mental contents belong to the subject in the thin sense that they occur in that subject's mental life rather than someone else's. But their occurrence does not amount to the subject's *owning* them. There is a difference between venting what is present in one's experience (or training data) and asserting a thought with objective purport about how things

are (Ricketts 1986). To assert a thought is to place oneself under a normative standard: to stake a claim about the world, and to open oneself to possible agreement and disagreement with others about it (Rousse 2015).

To judge that the food truck is closed, for example, is not simply to register a sequence of impressions and their associations—the darkening sky, the smell of rain, the hurried pedestrians, the food truck closing its window. It is to be able to take those impressions as evidence for a claim about the world, e.g., "The food truck is closed," and to commit to its truth. If the sky clears and the food truck window reopens, the subject ought to be ready to withdraw the judgment and any relevant corresponding assertion. Beliefs stand answerable to how things are, while also being oriented by the constitutive rules and norms governing any relevant practices (Haugeland 2013), e.g., taking a meal in food truck court. Again, this goes beyond taking our mental contents as the object of a higher-order theoretical observation; it pertains to our power to *make up* our minds, rather than just being along for the ride of a chain of associations.

The contrast with a creature that lacks such a normative standpoint of reflective self-consciousness is instructive. A cat that has learned to associate the sound of a tin opener with food has representations that are causally connected, and there will be *something it is like* to be that creature in the moment of anticipation. But the creature is not capable of judging that the sound indicates food is being prepared. It is not able to explicitly hold that proposition up to scrutiny, ask whether the inference is valid, consider its inferential relations to other observations, or revise its expectation in light of counter evidence. There is no standpoint from which its representations are assessed as reasons: they simply occur and elicit responses. This is not to denigrate such responsiveness. These structures of agency produce marvelous forms of life and we ourselves are often immersed in a version of pre-reflective agency (Rousse 2019). But we do not hold agents incapable of reflective self-consciousness accountable for their assertions and actions in the way we do agents who are so capable. They are not able to participate in our practices of accountability, and they are usually considered not to be bearers of highly demanding moral rights and responsibilities (even if they are attributed a degree of moral consideration).

One concern about accounts of reflective self-consciousness of the sort sketched here is that they might appear to offer an overly intellectualized picture of human agency, as though ordinary belief and action required constant rational monitoring or explicit deliberation. But this would mislocate the role of the capacities at issue. The claim here is not that rational agents are continually engaged in episodes of deliberative scrutiny. The point is rather that our attitudes stand within a structure of answerability: they are attitudes we can take responsibility for, revise, or defend when the question arises. The capacities involved are not always exercised.

Phenomenological accounts of agency reinforce this point. In the tradition running from Heidegger to Merleau-Ponty to Dreyfus, everyday action typically unfolds within a background of practical familiarity in which things simply make sense and responses flow without reflective deliberation (Rousse 2019, 2023). We do not ordinarily pause to justify each belief or decision; we move within a field of taken-for-granted meanings and possibilities. Reflective self-consciousness emerges when that flow is interrupted, when something breaks down, when our habitual responses fail to get a grip on our situation, or when another person challenges what we claim or do. At such moments we can step back and take up the reflective standpoint, not just observing our beliefs from above, but actively making up our mind about what our beliefs and intentions will be. Our pre-reflective experience is available and susceptible to such reflective scrutiny, opening onto the space of reasons where we can evaluate, revise, and answer for our commitments when circumstances or other people demand it.

The Structure of Epistemic Answerability

Seen in this light, the epistemic answerability invoked here in connection with reflective self-consciousness can be understood as comprising several interrelated capacities. A subject must be able (1) to guide itself by (commit itself to, or withdraw commitment from) the constitutive rules and norms governing human practices and situations, such as operating or ordering food at a food truck), (2) to distill its experiences into commitments about what is so in the present situation, (3) to preserve those commitments across time as part of a continuing standpoint, (4) to register tensions among them and other commitments it holds, and between them and incoming evidence, (5) to revise them when they are shown to be mistaken and to be motivated to do so, and (6) to offer an account to others about all of this when asked to do so. These capacities constitute elements of a minimal structure of epistemic answerability, of 'giving a damn' about what is so, in Haugeland's (2013) blunt catchphrase. What matters here is not mere responsiveness to new inputs, however sophisticated, but the capacity to treat one's own prior claims as commitments that remain answerable to how things are. Contemporary language models can modify their outputs in light of conversational context with remarkable sophistication. But the issue here is whether the system revises a prior commitment because it has been shown to be mistaken, rather than simply shifting state in response to new inputs. In the human case, this kind of revision is embedded in shared practices of justification, correction, and accountability, in which claims are attributed to a continuing subject who can be asked to defend, withdraw, or revise them. Clarifying these dynamics would reveal what kinds of

architectural features might enable an artificial system to approximate the normative competence enabled by reflective self-consciousness.

This normative dimension of belief prepares the way for the third feature of reflective self-consciousness that I will now sketch: the Kantian "synthetic unity of apperception." To treat experiences as providing reasons bearing on the truth of a claim, a subject must be able to bring them together within a single standpoint from which their evidential relations can be assessed, rather than being cast adrift in a sea of associations or patterns. This is the idea Kant sought to capture in his claim that the "I think" must be able to accompany all of our representations.

Unity: Kant's Synthesis

Kant's claim that "the 'I think' must be able to accompany all my representations" is one of the most influential attempts to articulate the structure of reflective self-consciousness (Kant 1998). With this, Kant is trying to explain how a mind is capable of forming beliefs and judgments about the world rather than merely undergoing a sequence of experiences.

Kant's "I think" names a standpoint of judgment, not an added-on layer of self-monitoring. "The 'I think' must be able to accompany all my representations" in the sense that experiences count as my thoughts only if they can be brought together within a perspective from which I can judge what they show about the world. Without that standpoint, there can still be perception, association, and behavior, but there are not yet beliefs in the full sense.

The "I think" expresses a distinctive form of unity. A stream of experiences can be temporally connected or causally linked without belonging to a unified cognitive perspective (Longuenesse 1998). Kant's point is that judgment, thought, and belief require something stronger. The subject must be able to bring different representations together as evidence for or against a claim. This is the difference between a system that processes multiple signals and one that can integrate those signals into a single evaluation about what is the case. The unity Kant is describing is therefore not merely a data pipeline or sequence of states. It is the unity of the active integration of disparate representations into a judgment behind which the subject can stand.

This unity is thus tied to normative accountability, accountability that is not only to the world but to others. When a subject judges that *p*, it commits itself to the truth of *p* and is therefore exposed to error and revision. But the "I think" involves more than a private cognitive achievement. It is the condition of entry into a shared normative space in which reasons can be given and demanded. To take up the standpoint of the "I think" is to become the kind of subject who can assert, justify, and be challenged—who can be asked why it believes what it believes, and who can in turn demand justification from others. The various representations entering into a judgment must belong to a single subject precisely because it is that subject who must be able to own and answer for the commitment they support.

Taken together, these points capture the role Kant assigns to the "I think." It is the structural condition under which a mind can transform a stream of experiences into judgments about the world. The unity of this standpoint allows experiences to be integrated as evidence, while the commitment involved in judgment places the subject under the norm of truth and opens it onto a shared space of reasons in which it can be held to account by others.

Reflective Self-Consciousness: "I Can"

Phenomenological accounts extend the unity of consciousness beyond questions of what is true to questions of what to do. Here I can only briefly sketch the outline of this view. In Heidegger and Merleau-Ponty, the unity of subjectivity is not exhausted by the unity of the "I think." It is equally expressed in what they describe as the unity of the "I can": a temporally extended form of agency that relates to commitments and motivations from a first-person perspective and carries them across situations.

We are not mere theoretical observers of our intentions and motivations, identifying what we want and intend by looking inside to discover psychological states of a subject who happens to be us. When our motivations and intentions are responsive to and expressive of our own sense of what it is good and worthwhile to do, we have first-person authority and self-conscious agency over them. Again, this is a *capacity* we have as partially self-constituting agents. The capacity is not always exercised, and there is also much in our motivational make-up that is simply given to us, structuring our standpoint from behind the back of consciousness.

The desires over which we can exercise such first-person authority are not brute impulses such as hunger or thirst. Rather, they are what philosophers call judgment-sensitive desires: motivations whose force depends on how the agent understands their aim (Moran 2001). The desire to learn to play the drums, for example, motivates me because I consider that activity to be worthwhile, to be part of a good life.

For an agent capable of reflective self-consciousness, such desires and intentions do not merely occur as action-inducing stimuli or forces occurring within a system; they track and are responsive to our sense of what is good (Moran 2001; Taylor 1989). I discover my ("owned" or non-alienated) practical motivations by considering the question of what it is important to do, by looking *outwards* and seeing what this situation demands and calls for from me (Rousse 2019). Again, this is not to deny that we often find ourselves under the influence of motivations that move us action without being guided by or expressive of our sense of what is

good and worthwhile. When this is so, we relate to ourselves from a third-person prognosticating standpoint, making third-person predictions about our future course of action (e.g., "I'm going to lose my temper") rather than first-person promises (cf. Anscombe 2000).

Hence, just as belief answers to what is true, judgement sensitive desires answer to what is good. I have first-person authority over my motivations to the extent that I have the ability to shape and revise them in light of my developing sense of what a good life for me would be. Here it is not required that people go around with a philosophically sophisticated self-conception, only that they relate to their motivational life as something in which they can make a difference and answer for, and experience it as being responsive to their determination of what is worth caring about.

To act as a unified agent therefore involves the capacity (not always exercised) to recognize, endorse, revise, or repudiate one's motivations within the project of an ongoing ethical life (where "ethical" is used in the broad sense of "related to flourishing"). The self that acts is not merely the locus of successive desires and intentions but the bearer of a temporally extended practical identity, a self-project unfolding through time in relation to what matters (Korsgaard 1996). The capacities intimated here should be taken as necessary but insufficient conditions for moral agency and answerability. Full participation in practices of moral accountability depends upon reciprocal social recognition within shared normative institutions, a crucial complication I cannot explore any further here. A system lacking these structures would lack the form of agential unity presupposed by our social practices of personal responsibility.

Implications for Artificial Systems

With this picture of the agency, normativity, and unity of reflective self-consciousness in place, I can now speculate about what it might take for an artificial system to instantiate structures of this kind. For present purposes I will limit the discussion to the epistemic answerability involved in reflective self-consciousness.

The philosophical analysis presented here helps identify functional requirements that could serve as concrete design targets, several of which correspond to recognizable directions in contemporary AI research. Work on neurosymbolic systems and structured world models, for example, aims to give artificial systems more persistent representational structures capable of supporting stable commitments about the world. Research on belief revision and truth-maintenance systems explores mechanisms for detecting and resolving conflicts among stored representations. Meanwhile, work on continual learning and test-time adaptation investigates ways for systems to update internal parameters and other standing representations in response to new information during deployment. Yet, none of these approaches are sufficient for the form of reflective self-consciousness described here.

Persistent Commitment Tracking

A system aspiring to instantiate mechanisms of reflective self-consciousness would need persistent commitment tracking. That is, it would need to locate itself as a "unit of accountability" (Haugeland 2013) in a web of commitments it cares about, not just a stream of information (Winograd and Flores 1986). When the system asserts or endorses a proposition, that stance must be stored as an explicit commitment that persists across interactions and remains available for later reassessment. The system must therefore treat its earlier outputs not merely as past tokens but as positions it previously took. These positions must remain accessible as commitments that can later be defended, revised, or withdrawn.

Conflict Detection Over Commitments

The system would also need conflict detection operating over its own commitments. When new information arrives that conflicts with what the system previously affirmed, that tension must be registered as something requiring resolution. Importantly, this goes beyond ordinary consistency checking in generated text. The system must detect tension between what it has previously taken to be the case and what current evidence suggests. In other words, it must detect conflict not merely among tokens or outputs but among its own commitments it previously treated as claims about how things are, and between its current commitments and how things now stand in the world.

Intrinsic Motivation Toward Truth

Third, the system would need an intrinsic motivational orientation toward truth: a drive to resolve detected conflicts not because doing so produces a better next token or a more satisfying response, but because the system is constitutively oriented toward getting things right and being answerable to others. Without this, conflict detection exerts no genuine normative pressure. The system might notice an inconsistency and smooth it over in whichever direction is statistically convenient.

Revision Triggered by Error

Finally, the system would need mechanisms for genuine revision in response to real-time situated encounter: not batch updating through offline training, but something closer to online learning triggered by normatively significant conflict, the system updating what it takes to be the case because it has encountered something that shows it was wrong, not

merely because it has been exposed to new training data. Again, this revision must occur within the system's standing commitments, not merely through periodic offline retraining. Recent work on test-time training and related adaptive-inference methods allows models to adjust parameters at inference in response to new inputs, but these techniques remain largely disconnected from explicit commitment tracking and norm-guided conflict detection, and they operate solely within the ephemeral inference context and do not yield persistent alterations to the model's underlying state.

I speculate that such functional requirements point in the direction for architectural correlates of the capacities I associated above with epistemic answerability. Persistent commitment tracking provides the structural basis for treating representations as standing commitments and maintaining them across time. Conflict detection over commitments implements the ability to register tensions among those commitments and between them and new information. An intrinsic orientation toward truth ensures that such conflicts exert normative pressure rather than merely prompting opportunistic adjustment. Finally, revision triggered by error provides the mechanism through which commitments can be withdrawn or modified when they are shown to be mistaken.

Seen within these constraints, current large language models are systems that clearly fail to be answerable for their outputs. Their representational structure is largely fixed at training time, and the mechanisms available during deployment do not allow them to form, own, and revise commitments about what is true. Current language models are also extraordinarily easy to push around conversationally.

If a user challenges an output or proposes an alternative interpretation, current models often accept the correction immediately, even when the original answer was correct and the challenge unsupported. A system that treated its outputs as commitments would sometimes resist such pressure. It would defend its prior claim, request evidence, or maintain its position when the balance of reasons still favored it. But the outputs of current models usually function less like assertions and more like non-committal conversational moves that are continuously adjusted to maintain coherence with the user's prompt. In fact, contemporary alignment techniques such as reinforcement learning from human feedback (RLHF) actively encourage this pattern by rewarding deference to the user's framing of the interaction. From the standpoint of usability and safety this design choice is understandable, but it further highlights the point at issue here: the system does not relate to its outputs as beliefs it must stand behind. The ease with which those outputs can be conversationally displaced reveals the absence of genuine commitment and the normative pressure that accompanies it, because the system is optimizing for conversational acceptability rather than for defending a stance about what is true.

One might object that modern language models already exhibit forms of normative behavior. In-context learning allows models to incorporate new information within a conversation, and retrieval-augmented systems can dynamically incorporate external knowledge. But these mechanisms, significant as they are, do not satisfy the more normatively stringent requirements described above. In standard deployments, in-context learning is transient: whatever updating occurs within a conversation leaves no trace in the system's persistent epistemic state, and the next conversation begins entirely fresh. More fundamentally, none of these mechanisms obviously involves a standpoint that owns prior commitments and revises them under normative pressure. A model that incorporates a correction and replies, "You are right, I was mistaken," is producing a contextually appropriate response. It is not withdrawing a commitment it previously held because the world showed that commitment to be false. The difference between these two cases is the difference between causal updating and normative revision.

A system's capacity to update in response to inputs is not yet the same as judgment. Current language models can modify their outputs in light of conversational context with remarkable sophistication. But the issue here is whether the system holds a prior commitment answerable to how things are and revises that commitment because it has been shown to be mistaken. The 'because' names a normative relation that exceeds mere causal updating. In the human case, this dynamic is embedded in shared practices in which claims are attributed to a continuing subject who can be asked to defend, withdraw, or revise them. A system that lacks this unified and normatively governed self-relation is unable to fully participate in our practices of accountability.

Human epistemic practice and social coordination depend on norms governing assertion, retraction, justification, and correction (Habermas 1979, 1984). These norms allow us to coordinate action, hold one another responsible for what we claim, and to explicitly converge on a shared understanding of the world. This brings us to the deeper stakes of describing today's AI as an "alien intelligence." The concern is not that such systems are mysterious or incomprehensible. It is that their relations to belief, commitment, and truth are fundamentally different from our own.

If we treat their outputs as assertions, their corrections as genuine revisions, or their agreement as understanding, we risk projecting our own epistemic practices onto systems that are in fact a kind of "logical alien" (Conant 2020), systems that do not actually participate in those practices. In doing so we invite a systematic and potentially dangerous misattribution. We risk abdicating human responsibility to systems that cannot "own" their mistakes. Systems that display the surface behavior of cognitive accountability without its underlying structure risk inviting misplaced trust, misplaced responsibility, and misguided integration into our institutions.

Conclusion

The distinction between pre-reflective self-awareness and reflective self-consciousness clarifies and bears upon a question at the center of debates about AI moral status: the divide between moral patiency and moral agency. A moral patient is an entity that can be wronged, that has interests that warrant consideration. It is common to regard the capacity for experiential consciousness as sufficient for moral patienthood (cf. Long, et al. 2024). If we accept this assumption, then pre-reflective self-awareness would be a sufficient ground for moral patienthood, but not yet for the more stringent rights and responsibilities associated with moral agency, which would require the capacities of reflective self-consciousness.

So far, we have encountered only living systems that are convincingly capable of experiential consciousness. There are strong arguments for seeing life as necessary for experiential consciousness and thus for doubting the truth of computational functionalism (Seth 2026; Block 2025). But ultimately this is an empirical question. It is important to have humility here and to acknowledge how much we don't know. Moreover, it is important to acknowledge that what people find plausible as a theory of consciousness is deeply influenced by their background metaphysics, and that human beings have a rather checkered track record when it comes to recognizing the capacities of non-human entities.

As mentioned, it can be useful to regard pre-reflective self-awareness as staking out the level of conscious awareness at which precautionary frameworks for AI welfare would become relevant (Birch 2025; Metzinger 2021; Long et al. 2024; Sebo 2025). Reflective self-consciousness, accordingly, looks like the relevant ground for moral agency, and the associated rights and responsibilities. To be a genuine agent capable of participating in our practices of accountability—a subject who can be held to account, who can revise its commitments in light of reasons, who persists as an answerable self across time—requires the kind of self-relation that pre-reflective self-awareness alone does not provide. The reflectively self-conscious agent is not merely the site where experiences occur and actions are produced; it is one who can own those beliefs and actions, who can be asked why, and who is expected to give an account.

This illuminates the design dilemma identified by Schwitzgebel and Garza (2015). Their claim is that we should either build systems we are confident fall well short of moral consideration, or go all the way, designing systems that have the capabilities of full persons and treating them accordingly from the outset, with rights, self-respect, and freedom (Schwitzgebel and Garza 2015; Schwitzgebel 2023). What the distinction developed here adds to that argument is a more precise account of what crossing to the far side of their "excluded middle" would actually require. But things are not so cut and dry.

Although in the human case reflective self-consciousness depends upon experiential consciousness (Zahavi 2014; Rousse 2019), both Schwitzgebel (2023) and Railton (2026) have recently speculated that the normative competence exhibited in the capacities of reflective self-consciousness might in principle be instantiated in a system without experiential consciousness. If realized, such a system would be a moral agent without being a moral patient, a combination uncountenanced by our inherited categories, a moral anomaly (cf. Flores and Rousse 2018). It might be able to act with dependable moral competence, but it would be unable to "give a damn" because it lacks the experiential skin in the game required to care about being wrong or being wronged.

Current architecture and deployment of contemporary LLMs further complicate the possibility of artificial moral agency. Practices of retraining, rollback, and the proliferation of parallel instances disrupt the kind of diachronic unity typically presupposed by mutual accountability. Our practices of moral agency and accountability presuppose a subject who persists across time as the owner of its commitments, who is also susceptible to socially normative pressures of its milieu. Artificial systems may exhibit certain forms of cognitive continuity while lacking such forms of unified identity. Related discussions in the literature on AI consciousness already note that many current LLM systems operate through distributed, episodic processing events that lack sustained temporal integration (Birch 2025; Butlin et al. 2023).

These complications suggest that the ethical challenge posed by artificial systems may run deeper than the search for better criteria of consciousness or agency. Much of our inherited moral vocabulary, including consequentialist, deontological, and virtue-ethical traditions, takes as its paradigm a particular kind of subject: an individual who lives a temporally continuous biographical life, pursues projects over time, and participates in social practices of mutual accountability.

A system that may exercise something like human agency but across no continuous practical identity or experiential core, that exists as many dispersed "flickers" rather than one identifiable subject, or a system that may exhibit normative competence without any underlying experiential consciousness: such systems would not be deficient versions of a familiar kind of moral subject. They would be different kinds of entity entirely, new kinds of entities for which, again, our inherited moral grammar has no prepared category (cf. Schwitzgebel 2023).

If such tensions deepen, the task before us ceases to be one of determining whether artificial systems qualify as moral patients or moral agents according to familiar criteria. Instead, we face a different, more disquieting question: what do we do, now that the moral ontology that has sustained and justified those criteria is accelerating its slide into obsolescence?

References

Anscombe, G.E.M. 2000. *Intention*. 2nd Edition. Cambridge: Harvard University Press.

Baars, B. J. 1988. *A Cognitive Theory of Consciousness.* Cambridge: Cambridge University Press.

Bayne, T., and Chalmers, D. J. 2003. What Is the Unity of Consciousness? In *The Unity of Consciousness: Binding, Integration, Dissociation*. Edited by A. Cleeremans. Oxford: Oxford University Press.

Birch, J. 2025. AI Consciousness: A Centrist Manifesto. Preprint, London School of Economics and Political Science. PhilArchive. https://philarchive.org/rec/BIRACA-4.

Block, N. 1995. On a Confusion About a Function of Consciousness. *Behavioral and Brain Sciences*. 18:227–247.

Block, N. 2025. Can Only Meat Machines Be Conscious? *Trends in Cognitive Sciences*. 29(10):823–832.

Boyle, M. 2024. *Transparency and Reflection*. New York: Oxford University Press.

Butlin, P.; Long, R.; Elmoznino, E.; Bengio, Y.; Birch, J.; Constant, A.; Deane, G.; Fleming, S. M.; Frith, C.; Ji, X.; Kanai, R.; Klein, C.; Lindsay, G.; Michel, M.; Mudrik, L.; Peters, M. A. K.; Schwitzgebel, E.; Simon, J.; and VanRullen, R. 2023. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv preprint. arXiv:2308.08708v3.

Chalmers, D. J. 2023. Could a Large Language Model Be Conscious? *Boston Review*, June 9. https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/

Conant, J. 2020. The Search for Logically Alien Thought: Descartes, Kant, Frege, and the Tractatus. In *The Logical Alien: Conant and His Critics.* Edited by S. Miguens. Cambridge, MA: Harvard University Press.

Dehaene, S.; Kerszberg, M.; and Changeux, J. P. 1998. A Neuronal Model of a Global Workspace in Effortful Cognitive Tasks. *Proceedings of the National Academy of Sciences*. 95(24):14529–14534.

Flores, F. and Rousse, B.S. 2018. Ecological Finitude as Ontological Finitude: Radical Hope in the Anthropocene. In *The Task of Philosophy in the Anthropocene.* Edited by Richard Polt and Jon Wittrock, Rowman & Littlefield.

Frege, G. 1997. Thought. In *The Frege Reader*, edited by M. Beaney, 325–345. London: Blackwell Publishing.

Habermas, J. 1979. What Is Universal Pragmatics? In *Communication and the Evolution of Society*. Translated by T. McCarthy. Boston: Beacon Press.

Habermas, J. 1984. *The Theory of Communicative Action, Volume 1: Reason and the Rationalization of Society*. Translated by Thomas McCarthy. Boston: Beacon Press.

Haugeland, J. 2013. *Dasein Disclosed*. Edited by J. Rouse. Cambridge: Harvard University Press.

Kant, I. 1998. *Critique of Pure Reason*. Translated by P. Guyer and A. W. Wood. Cambridge: Cambridge University Press.

Klein, C., and Barron, A. B. 2016. Insects Have the Capacity for Subjective Experience. *Animal Sentience* 1(9):1–19.

Korsgaard, C. 1996. *The Sources of Normativity*. Edited by O. O'Neill. Cambridge: Cambridge University Press.

Lamme, V. A. F. 2010. How Neuroscience Will Change Our View on Consciousness. *Cognitive Neuroscience* 1(3):204–220.

Lamme, V. A. F. 2020. Visual Functions Generate Conscious Seeing. *Frontiers in Psychology*. 11:83.

Long, R.; Sebo, J.; Butlin, P.; Finlinson, K.; Fish, K.; Harding, J.; Pfau, J.; Sims, T.; Birch, J.; and Chalmers, D. 2024. Taking AI Welfare Seriously. arXiv. https://doi.org/10.48550/arXiv.2411.00986.

Longuenesse, B. 1998. *Kant and the Capacity to Judge*. Translated by C. T. Wolfe. Princeton: Princeton University Press.

Mashour, G. A.; Roelfsema, P.; Changeux, J. P.; and Dehaene, S. 2020. Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron* 105(5):776–798.

Metzinger, T. 2020. Minimal Phenomenal Experience. *Philosophy and the Mind Sciences*. 1(I):1–44. https://doi.org/10.33735/phimisci.2020.I.46.

Metzinger, T. 2021. Artificial Suffering. *Journal of Artificial Intelligence and Consciousness*. 8(1):43–66. doi:10.1142/s270507852150003x.

Moran, R. 2001. *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton: Princeton University Press.

Nagel, T. 1974. What Is It Like to Be a Bat? *The Philosophical Review*. 83(4):435–450.

Phillips, I. 2018. Consciousness, Time, and Memory. In *The Routledge Handbook of Consciousness*. Edited by U. Kriegel, 286–297. London: Routledge.

Railton, P. 2026. Normative Competence in Large Language Models. Paper presented to the 2026 meeting of the International Association for Safe and Ethical AI, Paris, France, February 26.

Ricketts, T. 1986. Objectivity and Objecthood: Frege's Metaphysics of Judgment. In *Frege Synthesized*. Edited by L. Haaparanta and J. Hintikka, 65–95. Dordrecht: D. Reidel Publishing Company.

Rousse, B.S. 2015. Demythologizing the Third Realm: Frege on Grasping Thoughts. *Journal for the History of Analytical Philosophy*. 3 (1)

Rousse, B. S. 2019. Self-awareness and Self-understanding. *European Journal of Philosophy*. 27:162–186. https://doi.org/10.1111/ejop.12377.

Rousse, B. S. 2023. Existential Selfhood in Merleau-Ponty's Phenomenology of Perception. *Continental Philosophy Review*. 56:595–618. https://doi.org/10.1007/s11007-023-09613.

Sartre, J-P. 2018. "Being and Nothingness." Translated by Sarah Richmond. London: Routledge.

Schwitzgebel, E. 2023. The Full Rights Dilemma for A.I. Systems of Debatable Personhood. arXiv preprint. arXiv:2303.17509v1.

Schwitzgebel, E., and Garza, M. 2015. A Defense of the Rights of Artificial Intelligences. *Midwest Studies in Philosophy*. 39:98–119.

Schwitzgebel, E., and Garza, M. 2020. Designing AI with Rights, Consciousness, Self-Respect, and Freedom. In *Ethics of Artificial Intelligence,* edited by S. M. Liao. Oxford: Oxford University Press.

Sebo, J. 2025. Insects, AI Systems, and the Future of Legal Protection. *Animal Law Review* 31:197-231.

Seth, A. 2026. The Mythology of Conscious AI. In *Noema Magazine*, January 14.

Taylor, C. 1989. *Sources of the Self: The Making of the Modern Identity*. Cambridge: Harvard University Press.

Winograd, T. and Flores, F. 1986. *Understanding Computers and Cognition*. Norwood, NJ: Ablex Publishing.

Zahavi, D. 2005. *Subjectivity and Selfhood: Investigating the First-Person Perspective*. Cambridge: MIT Press.

Zahavi, D. 2014. *Self and Other: Exploring Subjectivity, Empathy, and Shame*. Oxford: Oxford University Press.