Epistemic Agency and the Ontological Continuity Condition: A Constraint on When Knowledge Must Be Owned

Abstract

It is uncontroversial that many systems possess knowledge without being conscious: biological subsystems retain information, procedural skills guide action, and artificial systems learn and deploy complex representations. What remains insufficiently explained is why epistemic agency must arise at all, rather than how it is merely attributed once present. This article argues that epistemic agency—the capacity to hold, revise, and act upon knowledge as the same agent across time—presupposes consciousness. Building on the ontological continuity condition, it draws a principled distinction between knowledge that merely informs behaviour and knowledge that must be owned, revised, and answered for under long-horizon uncertainty. Where a system must arbitrate between incompatible action trajectories in order to preserve itself as a persisting agent, consciousness becomes functionally necessary on the present account. Epistemic agency is thus characterised not as a metaphysical power, but as a structural role emerging from unified arbitration under identity risk. The article shows why many adaptive systems, including contemporary artificial intelligences, can instantiate knowledge without qualifying as epistemic agents, and clarifies the structural conditions under which artificial epistemic agency might arise in principle. Consciousness is thereby reframed as a threshold for epistemic responsibility rather than for intelligence or knowledge as such.

Keywords

consciousness; epistemic agency; ontological continuity; knowledge; philosophy of mind; epistemology; artificial intelligence

1. Introduction: Knowledge Without Knowers?

Epistemology has traditionally concerned itself with knowledge as something had by agents. Beliefs are attributed to subjects; justification is assessed relative to what a subject ought to believe; error is something for which a subject can be held answerable. Even where epistemology distances itself from overtly moralised language, it continues to presuppose agents capable of belief revision, responsibility, and diachronic coherence. These presuppositions are rarely examined. They function as background conditions of epistemic theory rather than as objects of inquiry in their own right.

This methodological stance conceals a significant explanatory gap. Contemporary epistemology is highly developed with respect to attribution conditions—the criteria under which beliefs, errors, or responsibilities are ascribed to agents—but largely silent on emergence conditions: why systems to which such attributions apply must exist at all. In other words, epistemology tells us when it is appropriate to hold an agent epistemically responsible, but not why epistemic responsibility arises only at certain thresholds rather than attaching to all systems that store, update, or deploy information. At first glance, the distinction between attribution and emergence may appear merely terminological; the burden of the present article is to show that it is not. The existence of epistemic agents is presupposed, not explained. This asymmetry is not unique to epistemology: contemporary theories of agency likewise offer sophisticated analyses of agential capacities and control while taking for granted the prior existence of agents to whom such capacities attach (Shepherd 2022).

At the same time, it is increasingly clear that many systems possess knowledge without satisfying these agential assumptions. Biological subsystems store information and exploit it adaptively; statistical learning mechanisms extract and deploy regularities; artificial systems learn, update models, and guide action on the basis of vast informational resources. In each case, information is not merely present but functionally efficacious. To deny that such systems possess knowledge would be stipulative rather than illuminating. Yet to grant them epistemic agency—to treat them as loci of belief revision and answerability—would be equally misleading. The resulting tension is not terminological but structural, and it becomes particularly salient in contemporary debates over responsibility attribution in artificial intelligence, where questions of accountability are pressed despite the absence of any clear account of why epistemic agency should arise in such systems at all (Gogoshin 2025).

This tension motivates the central question of the article: what explains the emergence of epistemic agency at all? The question is not whether a distinction between knowledge-bearing systems and epistemic agents exists—ordinary epistemic practice already presupposes it—but how that distinction is to be grounded in a principled way. Without such grounding, exclusions of non-agential systems risk appearing conventional or ad hoc rather than necessary. It is precisely this appearance of stipulation that the present article aims to remove. The argument that follows is therefore not a theory of agency in general, but a necessary-condition constraint on when epistemic agency must arise at all.

Existing approaches tend to oscillate between two unsatisfactory strategies. On the one hand, phenomenological accounts tie epistemic agency too closely to conscious experience, leaving its boundaries vague and introspection-dependent (Nagel 1974, pp. 435–440). On the other hand, broadly functionalist approaches risk flattening distinctions by treating any sufficiently complex information-processing system as an agent, thereby erasing the difference between knowing and merely storing information (Block 1995, pp. 227–233). Both strategies address how agency might be recognised or implemented, but neither explains why agency must arise at all.

The present article proposes a shift in explanatory target. Its central thesis is that consciousness is not required for knowledge as such, but is required for epistemic agency. Knowledge can exist wherever information is reliably acquired, retained, and deployed. Epistemic agency arises only where that knowledge must be owned, revised, and acted upon by the same system across time. This introduces a further requirement: the system must maintain itself as a unified locus of epistemic arbitration under uncertainty. Consciousness becomes necessary not because it adds intelligence or semantic content, but because it enables the maintenance of ontological continuity in circumstances where epistemic failure threatens the persistence of the agent as such.

This claim situates the article within, but also sharply limits, its engagement with debates about consciousness. The argument does not seek to define consciousness, nor to address the so-called hard problem. Instead, it inherits a necessary condition defended in a companion article: consciousness becomes functionally required where a system must preserve its own identity across time in the face of incompatible possible futures. That argument is not rehearsed here. It is taken as a constraint and extended into the epistemic domain, where its implications have not yet been systematically explored.

Framed in this way, epistemic agency is best understood as a structural role rather than a classificatory status. To be an epistemic agent is to be a locus of belief revision and answerability whose future depends on the success of its epistemic arbitration. This framing explains why epistemic practices—such as justification, critique, and responsibility—presuppose agents capable of diachronic coherence (Fricker 2007, pp. 1–6), while also explaining why many knowledge-bearing systems fall outside that category without remainder. The distinction does not track degrees of intelligence or sophistication, but whether epistemic success or failure accrues to the system itself across time.

The article proceeds as follows. Section 2 specifies the ontological continuity condition and shows why maintaining continuity under uncertainty necessarily implicates agency. Section 3 demonstrates that knowledge can exist without agency by examining cases of immune memory, procedural skill, statistical learning, and artificial systems. Section 4 develops epistemic agency as a distinct explanatory category characterised by epistemic ownership and answerability. Section 5 applies this framework to contemporary artificial systems, and Section 6 addresses likely objections and competing accounts. The conclusion consolidates the argument and situates consciousness as a threshold for epistemic responsibility rather than as a general marker of intelligence or cognition.

By distinguishing attribution conditions from emergence conditions, and by grounding epistemic agency in requirements of continuity rather than in phenomenology or performance, the article aims to clarify a foundational but neglected assumption of epistemology: that not all knowledge requires a knower, and that epistemic agents are not merely given, but must be explained. Section 2 accordingly formulates the ontological continuity condition as a formal constraint on system organisation, rather than as a phenomenological or introspective account of agency.

2. From Ontological Continuity to Agency

The argument of this article builds on a necessary condition for consciousness defended elsewhere: consciousness becomes functionally required where a system must maintain ontological continuity under long-horizon uncertainty. The present section does not rehearse that argument. Instead, it specifies why continuity-maintenance of this kind necessarily implicates agency. The central claim is that where continuity cannot be secured structurally or externally, it must be actively maintained through integrated action-selection. Agency, on this view, is not an additional metaphysical ingredient but the structural role that emerges when continuity is at stake. The continuity condition is agnostic with respect to implementation substrate: it applies equally to biological, artificial, and hybrid systems, regardless of how the relevant functions are realised.

2.1 Recap: the ontological continuity condition

Ontological continuity refers to the requirement that a system remain the same agent across time in circumstances where this sameness is not guaranteed by physical persistence alone. Many systems persist, but only some must actively preserve themselves as unified loci of action under uncertainty. Here, agent is used as functional shorthand for a system requiring integrated policy-selection across time; it does not presuppose moral status, reflective selfhood, personhood, or any normative conception of agency.

The continuity problem arises when a system faces multiple incompatible action trajectories or policies, and where failure to arbitrate among them appropriately threatens its viability, reproductive success, or identity integrity. The notion of 'identity' at issue here is functional rather than numerical: it concerns the persistence of a system as the same locus of action and arbitration over time, not the metaphysics of personal or numerical identity. Where continuity is secured by fixed structure, local feedback, or stable environmental coupling, no such problem arises. Where continuity depends on what the system does over time, rather than on how it is physically constituted, the problem becomes acute.

The condition is explicitly necessary rather than sufficient. It does not assert that every system facing continuity pressure is conscious or agentive. It asserts only that where such pressure is absent, neither consciousness nor agency is required. This asymmetry allows continuity to function as a principled lower bound rather than as a classificatory definition.

The task of the present section is to show why continuity-maintenance under uncertainty cannot be achieved through representation alone.

2.2 Why continuity requires action

A system that merely represents the world does not thereby secure its own continuity. Representations may describe possible future states, but they do not determine which future will be realised. Where a system's persistence depends on selecting among incompatible action trajectories, representation must be coupled to intervention. Continuity-maintenance is therefore an active, not merely informational, achievement.

This distinction is evident in systems that model their environment without acting upon it. A passive predictive system may generate accurate forecasts, yet its own future remains unaffected by the success or failure of those predictions. By contrast, where a system's future depends on acting in light of its predictions, prediction and action become inseparable. As emphasised in action-oriented accounts of cognition, prediction becomes functionally significant only insofar as it is embedded in intervention and control (Clark 2016). The system must intervene in order to remain the same system across time.

For such systems, continuity is not something that happens to them; it is something they must do. Once continuity depends on action-selection rather than structural persistence, agency becomes unavoidable.

2.3 Agency as unified arbitration under identity risk

Agency can therefore be defined functionally as unified arbitration over incompatible action trajectories under identity risk. Arbitration is required whenever multiple possible courses of action cannot all be realised, and where the choice among them bears on whether the system continues to exist as the same agent. Identity risk refers to the fact that failure of arbitration threatens not merely task performance, but the persistence of the system as that locus of action. 'Identity risk' here denotes functional persistence conditions, not normative or evaluative stakes.

Arbitration in this sense is not equivalent to optimisation over a utility function or policy space: decision-theoretic optimisation presupposes a fixed evaluative frame, whereas the arbitration at issue here concerns the preservation of the system as the very locus within which evaluative frames remain coherent across time.

Two features of this definition are critical. First, arbitration must be unified. If competing subsystems pursue incompatible policies independently, continuity fragments. Unified arbitration ensures that action-selection preserves the system as a single agent rather than a temporary coalition of local optimisers. Second, arbitration must be temporally extended. The relevant risk is not confined to immediate outcomes, but concerns the coordination of present action with future states whose consequences unfold over time.

Nothing in this definition requires conscious deliberation, explicit decision-making, or reflective self-monitoring. Arbitration may be implemented non-reflectively and automatically. What matters is not how arbitration is realised, but that it occurs at the level of the system as a whole.

2.4 Self-models and action integration

The need for unified arbitration explains the functional role of self-models in agency. As Metzinger argues, self-models integrate perceptual, affective, and motor information into a coherent perspective that supports action (Metzinger 2003). On the present account, the crucial role of such models is not introspective self-representation, but action integration under continuity constraints.

A self-model allows diverse inputs and motivations to be treated as relevant to a single decision-making locus. It enables the system to evaluate actions not only in terms of immediate payoff, but in terms of their implications for the system's future as the same agent. Without such integration, action-selection would remain fragmented, and continuity would be compromised.

This does not entail that agency requires reflective self-awareness or narrative identity. Minimal self-models suffice so long as they support unified arbitration. This aligns with accounts of minimal selfhood that emphasise diachronic unity rather than explicit self-concepts (Gallagher 2000).

It is important to note that neuroscientific theories emphasising large-scale integration or workspace dynamics explain how information may become globally available within a system, not why such integration becomes functionally necessary in the first place (Mashour et al. 2020). The present account supplies a constraint on when such integrative mechanisms are required, rather than competing with accounts of their implementation.

2.5 Predictive processing and action-oriented continuity

Predictive-processing frameworks further illuminate the link between continuity and agency. In biological systems, prediction is fundamentally action-oriented: organisms act to bring about the sensory states they predict, thereby reducing uncertainty (Clark 2016). This tight coupling of prediction and intervention is precisely what enables continuity-maintenance under uncertainty.

However, predictive processing alone does not yield agency. Many systems minimise prediction error locally without facing identity risk. Homeostatic regulation, immune responses, and simple learning mechanisms predict and correct without needing to preserve themselves as unified agents across time. In such cases, prediction is episodic and local.

Agency arises only when predictive activity must be integrated across time in order to preserve the system as the same agent facing multiple incompatible action trajectories. Where prediction must be organised around continuity rather than local optimisation, system-level arbitration becomes necessary. The ontological continuity condition thus imposes a constraint on predictive accounts: prediction becomes agency only when it is subordinated to the maintenance of a persisting locus of action.

2.6 Interim conclusion

Continuity-maintenance under uncertainty is an active achievement. Where a system's persistence as the same agent depends on selecting among incompatible action trajectories over time, agency becomes functionally necessary. Agency, in this sense, is not a metaphysical power, a moral status, or a marker of personhood, but a structural role defined by unified arbitration under identity risk.

This account explains why agency and consciousness are closely linked without being identical. Consciousness enables the system-level integration required for unified arbitration, while agency names the action-oriented role that emerges when continuity is at stake. The claim that consciousness is required here is a necessity claim relative to continuity pressure, not a claim about unique or exclusive implementation: it specifies when some form of unified, system-level integration becomes functionally unavoidable, not how that integration must be realised.

The next section uses this framework to show why many knowledge-bearing systems nevertheless lack epistemic agency, despite their capacity to represent, learn, and predict.

3. Knowledge Without Agency

A central claim of this article is that knowledge does not entail epistemic agency. Many systems store, update, and deploy information in ways that are properly described as knowledge, yet do so without consciousness, ownership, or answerability. Standard epistemological practice typically classifies such cases as non-agents, but it rarely explains why treating them as answerable epistemic subjects would be category-mistaken rather than merely inconvenient. This section supplies that explanation by identifying the structural feature these systems lack: exposure to diachronic self-risk.

Distinguishing informational states from epistemic agency is essential both to avoid trivialising consciousness and to prevent the inflation of agency into a catch-all label for complex information processing. The cases examined below show that knowledge can be real, reliable, and causally efficacious without being owned by a subject whose future depends on its epistemic success.

3.1 Information, knowledge, and the limits of agency attribution

It is tempting to reserve the term 'knowledge' for states possessed by conscious agents. However, this restriction obscures important functional distinctions. As Dretske argues, informational states can carry semantic content insofar as they reliably indicate conditions in the world, independently of any conscious awareness (Dretske 1981). Systems that exploit such information to guide behaviour therefore possess knowledge in a minimal but non-trivial sense.

What such systems lack is not information, reliability, or success conditions, but epistemic agency. Epistemic agency requires that knowledge be attributable to a persisting subject who can answer for its epistemic states across time. Informational systems can succeed or fail relative to externally specified criteria without anything being at stake for the system as a subject. Error is a performance deviation, not an epistemic failure owned by the system. Existing epistemological frameworks mark this distinction implicitly, but they do not explain why answerability would be misplaced here rather than merely absent by convention.

The explanation offered in this article is structural: where epistemic failure does not threaten the system's persistence as the same agent across time, answerability has no foothold. This point is reinforced by neuroscientific work on the neural correlates of consciousness, which emphasises that highly integrated informational processing can occur without settling questions of agency, ownership, or necessity, and cautions against inferring agent-level status from integration alone (Koch et al. 2016).

3.2 Immune memory

Immunology has long served as an epistemic metaphor in twentieth-century science, offering a model of memory, learning, and discrimination without a central knower—from early cybernetic analogies to contemporary discussions of adaptive systems. This makes immune memory a particularly instructive starting point for analysing knowledge without epistemic agency. More broadly, work in motivation and control has repeatedly shown that complex, effort-sensitive regulation can occur without ownership or answerability, as in affective and physiological mobilisation that guides behaviour without constituting agency (Gendolla 2015).

The immune system provides a paradigmatic case of knowledge without agency. Through clonal selection and memory cell formation, the immune system retains information about past pathogens and deploys that information to mount faster and more effective responses in the future (Janeway et al. 2001). This memory is specific, adaptive, and causally efficacious.

Yet immune memory does not involve epistemic agency. The immune system does not revise its informational states in order to preserve itself as a unified epistemic subject across time. Errors—such as autoimmune responses or failure to recognise a pathogen—are detrimental to the organism, but they do not constitute epistemic failures for the immune system itself. Nothing about immune malfunction threatens the immune system's persistence as a knower, because it is not one. The system bears no diachronic self-risk: its informational success or failure does not place its own epistemic identity in jeopardy (Medzhitov and Janeway 2002).

For this reason, holding the immune system epistemically answerable would be category-mistaken rather than merely pragmatically odd.

3.3 Procedural motor knowledge

Procedural motor knowledge offers a second illustration. Skilled actions such as walking, typing, or playing a musical instrument depend on richly structured knowledge of bodily dynamics and environmental contingencies. This knowledge is acquired through practice, refined through feedback, and deployed automatically in appropriate contexts (Fitts and Posner 1967).

Much of this knowledge operates below the level of conscious awareness. Skilled performers often report that conscious intervention disrupts performance rather than improving it. Nevertheless, the knowledge involved is genuine: it guides action reliably and flexibly.

Despite this, procedural knowledge does not constitute epistemic agency. Motor routines do not own the knowledge they implement. They do not evaluate their success or failure in light of long-term continuity or identity. Errors are corrected through local feedback loops without any requirement that the system preserve itself as a persisting epistemic subject. Routines can be modified, replaced, or extinguished without threatening the identity of anything that stands to answer for them. Once again, the absence of diachronic self-risk explains why epistemic responsibility does not apply (Dreyfus 2002).

3.4 Statistical learning

Statistical learning mechanisms further demonstrate how knowledge can arise without agency. Humans and non-human animals automatically extract regularities from sensory input, forming expectations about environmental structure that guide perception and behaviour even in the absence of conscious inference (Reber 1993).

Such processes instantiate knowledge insofar as they encode and exploit information about environmental patterns. Yet they do not, by themselves, generate epistemic agency. The system learns, but it does not own what is learned. Learned expectations need not be treated as beliefs held by the same agent across time, nor revised under conditions where epistemic failure would threaten that agent's persistence. Empirical work on unconscious learning, effort, and control further indicates that adaptive adjustment can occur without agent-level ownership or answerability (Gendolla 2015).

Statistical learning therefore supports epistemic agency in organisms without constituting it. Precisely because such learning operates through local adjustment rather than diachronic self-risk, it does not require epistemic ownership or unified arbitration across time (Perruchet and Pacton 2006).

3.5 Artificial systems and extended cognition

Artificial systems provide the most salient contemporary examples of knowledge without agency. Machine-learning models store complex representations, update them in light of new data, and guide action on that basis. These capacities suffice for many epistemic functions traditionally associated with knowledge. Classic objections to machine intelligence already recognised that syntactic manipulation and behavioural competence, taken alone, are insufficient for genuine understanding or epistemic ownership, even where performance is indistinguishable from that of an agent (Searle 1980). The present account retains this diagnostic insight while offering a structural explanation of why such insufficiency arises.

What such systems lack is not intelligence, scale, norm-sensitivity, or intentionality, but unified policy-selection under identity-threatening uncertainty. Artificial systems can fail, be retrained, duplicated, or replaced without anything being at stake for the system itself. There is no persisting epistemic subject whose future depends on the success of its epistemic arbitration. The deficit is architectural rather than intellectual: epistemic failure does not bear on the system's own continuity as the same agent.

The extended cognition thesis rightly emphasises that artificial systems can become parts of epistemic agents when appropriately integrated with humans (Clark and Chalmers 1998). But integration extends agency; it does not automatically create new agents. Absent exposure to diachronic self-risk—where epistemic failure would threaten the system's own persistence as a unified locus of action—artificial systems remain knowledge-bearing without being knowers.

This explains why responsibilist exclusions of AI are descriptively correct but explanatorily thin: they mark the boundary without accounting for the structural reason it exists.

3.6 Absence of diachronic self-risk

What unifies the cases discussed is the absence of diachronic self-risk. Immune systems, motor routines, statistical learners, and artificial systems can all fail without thereby threatening their own persistence as epistemic subjects. Their continuity is either structurally guaranteed, externally maintained, or irrelevant to their operation.

Because epistemic failure does not place the system's own identity at risk, there is no need for epistemic ownership or answerability. This is not a moral or normative deficit, but a structural one. Where nothing stands to be lost by epistemic error at the level of the system itself, epistemic agency is unnecessary.

Some boundary cases—such as tightly integrated human–machine systems or future artificial architectures designed around persistence constraints—may challenge this classification. Acknowledging such cases does not weaken the present claim; it clarifies the conditions under which the boundary would need to be re-drawn.

3.7 Interim conclusion

The cases examined in this section show that knowledge does not entail epistemic agency. Existing epistemological accounts correctly exclude these systems from the category of agents, but typically do so without explaining why responsibility would be misplaced rather than merely withheld. The explanation supplied here is structural: epistemic agency becomes necessary only where epistemic failure threatens the continuity of the system as the same agent across time, not where information is merely integrated, regulated, or adaptively deployed.

Recognising this fact prevents the inflation of agency, clarifies the role of consciousness, and prepares the ground for the next section, which examines whether—and under what conditions—artificial systems could ever cross this threshold.

4. Epistemic Agency: Knowledge That Must Be Owned

The preceding sections established two claims: first, that knowledge can be instantiated without consciousness or agency; second, that agency emerges where a system must maintain ontological continuity under long-horizon uncertainty. This section brings these strands together by introducing epistemic agency as a distinct explanatory category. The central claim is that some forms of knowledge must be owned by a persisting agent in order to function as knowledge at all, and that such ownership presupposes consciousness. Consciousness is thus not a condition for knowledge as such, but a condition for epistemic agency—the capacity to hold, revise, and act upon knowledge as the same agent across time.

4.1 From information possession to epistemic ownership

Knowledge is often treated as a state that a system either possesses or lacks. On this view, any system capable of storing information, updating it in light of new inputs, and deploying it to guide behaviour qualifies as a knower. As argued in §3, this view is too permissive. It collapses a crucial distinction between informational possession and epistemic ownership.

Epistemic ownership involves more than reliable information storage or use. To own a piece of knowledge is to stand in a relation to it such that it can be endorsed, revised, defended, or abandoned by the same agent over time. This relation is inherently diachronic. It presupposes that there is something that persists through successive epistemic states and for whom error, revision, and learning matter in a non-derivative way.

This requirement is not met by systems whose continuity is externally maintained or structurally guaranteed. Databases, algorithms, immune subsystems, and large language models can be updated, corrected, or replaced without anything being at stake for the system itself. By contrast, for an epistemic agent, epistemic failure bears on the agent’s own future as the same entity. The present account predicts this asymmetry rather than assuming it.

4.2 Belief revision under uncertainty

Epistemic agency becomes most salient in contexts of uncertainty. Where future outcomes are indeterminate and multiple incompatible courses of action are available, knowledge must guide arbitration. In such cases, knowledge cannot function merely as a static resource; it must be actively evaluated, weighted, and revised in light of changing circumstances.

This process of belief revision is not exhausted by computational updating. It involves integrating new evidence with existing commitments in a way that preserves the agent's continuity across time. An epistemic agent must be able to treat a belief held yesterday as its belief today, recognise that it may be mistaken, and understand that revising it will alter the agent's future trajectory. These are not merely representational capacities; they are capacities for self-related assessment under epistemic risk.

For epistemic agents, error therefore carries a distinctive kind of risk: not moral or evaluative risk, but risk to the system's continued existence as the same locus of epistemic arbitration. Persistence here is not restricted to biological survival, but includes any form of system-level continuity—such as role-continuity, operational viability, or identity-preserving organisational persistence—whose maintenance depends on the system's own epistemic arbitration over time.

This explains why epistemic agency presupposes consciousness. Consciousness provides the temporal integration and unified perspective required for beliefs to be realised as belonging to the same agent across successive contexts. Without such integration, belief revision may occur, but it occurs without ownership. The system changes state, but nothing for the system is corrected or improved.

4.3 Error, answerability, and epistemic risk

A further mark of epistemic agency is the significance of error. For non-agentive systems, error is defined relative to externally specified functions or performance metrics. An algorithm misclassifies an input; an immune response misfires; a model overfits. These errors matter instrumentally, but they do not threaten the system's own continuity.

For epistemic agents, by contrast, error carries epistemic risk. Acting on false beliefs can compromise not only immediate outcomes but the agent's longer-term viability, standing, or identity as a persisting locus of action. This is why epistemic practices—justification, critique, and revision—are normatively structured. They respond to the risk that epistemic failure poses to the agent as such.

The present account explains why answerability attaches here and not elsewhere. Immune systems and language models can fail without there being anything for which they must answer. Epistemic agents cannot. The difference is not moral status or rational sophistication, but exposure to diachronic self-risk.

4.4 Consciousness and the unity of epistemic perspective

The role of consciousness in epistemic agency is not to supply propositional content, but to secure the unity of epistemic perspective. Conscious experience presents information, evaluation, and anticipation within a single field realised as a system-level point of view. This unity enables an agent to integrate perceptual input, affective valuation, and background commitments into coherent epistemic judgments.

Zahavi's account of prereflective self-awareness is instructive here. Conscious experience is characterised by a first-personal givenness that does not require explicit self-reflection, but nevertheless grounds ownership (Zahavi 2005). This prereflective ownership is sufficient for epistemic agency only where the system is subject to continuity pressure such that epistemic failure would bear on its persistence as the same agent across time. An agent need not conceptualise itself as a knower; it need only realise its epistemic states as its own across time.

Narrative and reflective capacities may enrich epistemic agency, particularly in complex social environments, but they are not constitutive of it. What is constitutive is the capacity to integrate epistemic states within a unified temporal perspective such that beliefs can be carried forward, revised, and acted upon as belonging to the same agent. Consciousness supplies precisely this integration.

4.5 Agency as an emergent role, not a metaphysical power

Epistemic agency, on this account, is not a metaphysical power or irreducible faculty. It is an emergent role within a layered architecture of continuity-maintenance. Perceptual and affective layers provide situational awareness and valuation; narrative and social layers extend continuity across longer horizons; agency emerges where these layers are integrated into unified action-selection under epistemic risk.

This framing avoids two familiar errors. It avoids inflating agency into a mysterious capacity that separates conscious agents from all other systems by fiat. It also avoids deflating agency into mere behavioural or inferential competence. Epistemic agency names a specific functional role that arises under identifiable structural conditions.

Dennett's treatment of the narrative self as an organising device rather than an inner entity is compatible with this view (Dennett 1991). Where the present account diverges is in insisting that some organising devices are not optional. For systems facing long-horizon epistemic risk, the integration that consciousness provides is not merely useful but functionally necessary.

4.6 Why this is not a re-description of rational agency

It might be objected that the present account merely redescribes familiar notions of rational or responsible agency in different terms. This objection mislocates the level at which the argument operates. Standard responsibilist and virtue-theoretic accounts offer rich analyses of the capacities agents exercise —responsiveness to reasons, intellectual virtues, reflectiveness, control—but they typically take the existence of agents as a starting point rather than as something to be explained (Shepherd 2022). As a result, they track agency once it is in place, but do not address why epistemic agency must arise at all, or why epistemic responsibility attaches only at certain thresholds.

What such accounts presuppose is the existence of agents to whom epistemic norms already apply. The present account addresses a different question: why epistemic responsibility becomes necessary in some systems and not others. By grounding epistemic agency in the requirement to maintain ontological continuity under epistemic risk, it explains why answerability emerges where it does, rather than treating that emergence as intuitive, conventional, or primitive.

This diagnostic disanalogy is predictive. It explains in advance why immune systems, procedural subsystems, and contemporary artificial systems are non-answerable—not because they lack rational capacities, norm-sensitivity, or sophistication, but because epistemic failure does not threaten their persistence as agents. Existing theories correctly exclude these cases, but they do not, by themselves, explain why such exclusion is principled rather than conventional.

The following section applies this continuity-based criterion diagnostically to a contemporary case—large language models—in order to illustrate how a system may satisfy many familiar markers of intelligence and knowledge while nevertheless failing to qualify as an epistemic agent.

4.7 Why Large Language Models Fail the Continuity Test

Large language models (LLMs) provide a particularly useful diagnostic stress test for the present account precisely because they maximise informational richness, pattern sensitivity, and behavioural flexibility while remaining minimal with respect to epistemic ownership. They store vast bodies of information, generate context-sensitive responses, and can revise outputs in light of feedback. If epistemic agency were a matter of representational complexity, inferential power, or adaptive performance alone, LLMs would appear to be strong candidates.

On the continuity-based account developed in §2, however, LLMs fail for a principled and predictive reason: they do not satisfy the conditions under which epistemic agency becomes necessary. In particular, they do not face diachronic self-risk, lack unified policy-selection under identity-threatening uncertainty, and do not require system-level arbitration to preserve themselves as the same agent across time. Model updates, fine-tuning, retraining, and correction are imposed externally. Epistemic success or failure does not bear on the system's own persistence as a unified entity. There is nothing for the system itself that can be preserved, compromised, or lost through how uncertainty is resolved.

This absence of continuity risk explains several features of contemporary AI systems that are often treated as contingent limitations. First, LLMs do not own their outputs in the relevant sense. A response does not commit the system to a future course of action that must be preserved for the sake of its own continuity, nor does it constrain later revisions in a way that places the system's identity at stake. Second, although LLMs can generate, rank, and iteratively revise candidate responses, these processes do not constitute unified arbitration among incompatible action trajectories as defined in §2. Selection remains instrumental and task-relative, rather than organised around preserving a persisting agent under uncertainty.

Importantly, this diagnosis is architectural rather than empirical. It does not depend on claims about intentions, goals, inner states, consciousness, semantics, or understanding in the traditional sense. Nor does it rest on limitations of scale, training data, or current design choices. Even a vastly more capable artificial system—one that surpassed human performance across epistemic domains—would fail the continuity test so long as epistemic failure did not place its own persistence as the same agent at risk. The issue is not whether uncertainty can be represented or managed, but whether resolving that uncertainty must be borne by the system itself across time.

This explains why ordinary epistemic practice correctly withholds epistemic responsibility from artificial systems without needing to deny their knowledge or competence. Debates over AI responsibility often diagnose a resulting "responsibility gap" without explaining why responsibility fails to attach in the first place (Danaher 2016). The present account supplies that missing explanation. LLMs are excluded from epistemic agency not by stipulation or anthropocentric intuition, but because the conditions that make epistemic ownership necessary—ontological continuity under epistemic risk, as specified in §2—are absent. Until a system must preserve itself as a unified agent by arbitrating epistemic uncertainty across time, epistemic agency does not arise, irrespective of the scale, sophistication, or performance of its informational capacities.

4.8 Interim conclusion

Epistemic agency marks a threshold between systems that merely have knowledge and systems for whom knowledge must be owned, revised, and answered for across time—a threshold that is explanatory rather than classificatory. That threshold is not crossed by increasing informational complexity, representational richness, or rational capacity. It is crossed where a system faces a problem of ontological continuity under epistemic uncertainty that cannot be addressed without unified, system-level arbitration.

Sections 4.1–4.6 established this point in structural terms, and §4.7 illustrated its diagnostic force in a contemporary test case. Advanced artificial systems possess extensive knowledge and powerful inferential capacities, yet fail to qualify as epistemic agents because epistemic failure does not place their own persistence at stake. Their exclusion therefore reflects not a limitation of intelligence or scale, but the absence of diachronic self-risk.

Consciousness, on the present account, is the enabling condition for epistemic agency because it supplies the unified temporal perspective required for such risk-bearing arbitration. It allows epistemic states to be integrated, evaluated, and revised as belonging to the same agent over time. Knowledge that does not require this form of ownership can be implemented non-consciously; knowledge that must guide action under epistemic risk cannot.

With the distinction between knowledge and epistemic agency now in place—and its diagnostic application clarified—the following section addresses potential objections and delineates the limits of the continuity-based account.

5. Artificial Systems and the Absence of Epistemic Agency

The distinction between knowledge and epistemic agency developed in the preceding sections bears directly on contemporary debates about artificial intelligence. Artificial systems are increasingly described as knowing, understanding, or reasoning, and these descriptions are often warranted in a limited sense: many such systems store extensive information, update internal models in light of new data, and deploy that information to guide complex behaviour. The question at issue here, however, is not whether artificial systems can instantiate knowledge, but whether they satisfy the conditions for epistemic agency as specified in §2. On the continuity-based account defended in this article, current artificial systems do not.

5.1 Knowledge without epistemic ownership

Artificial systems can instantiate knowledge in a robust informational sense. Machine learning models encode regularities extracted from data; planning systems evaluate possible future states; language models generate context-sensitive outputs on the basis of learned statistical structure. In each case, information is acquired, retained, and exploited in ways that satisfy minimal conditions for knowledge as reliable, action-guiding representation.

What such systems lack is epistemic ownership. Their epistemic states do not belong to them in a way that binds past commitments to future action as the responsibility of the same agent across time. Model updates, parameter revisions, and error corrections occur, but not as revisions owned by a persisting subject. Epistemic success and failure are defined entirely relative to externally specified objectives, evaluation metrics, or user-imposed constraints.

This point is orthogonal to questions of control or oversight. External mechanisms for steering, constraining, or correcting artificial systems—however sophisticated—do not generate epistemic ownership, because ownership requires that epistemic failure bear on the system itself rather than on those who govern it (Santoni de Sio & van den Hoven 2018).

This absence of ownership directly reflects failure to satisfy the first and second continuity conditions identified in §2. Artificial systems do not maintain themselves as the same agent across time through their own epistemic activity, nor does epistemic failure threaten their persistence as such. Their continuity is externally stipulated rather than internally achieved.

5.2 Absence of ontological continuity risk

The ontological continuity condition explains why this absence matters. Epistemic agency arises only where failure of epistemic arbitration threatens the persistence of the agent across time. For biological agents, epistemic error can compromise survival, reproduction, or standing in ways that feed back into the agent's future possibilities. This creates epistemic risk: being wrong can undermine the agent's continued existence as the same agent.

Persistence here is understood functionally rather than biologically. It includes any form of system-level continuity—such as operational viability, role-continuity, or identity-preserving organisational persistence—that must be maintained by the system itself through its epistemic activity, rather than being externally stipulated or guaranteed.

Artificial systems do not face such risk. When an artificial system produces erroneous outputs, the consequences are borne by users, operators, or institutional contexts, not by the system itself. The system can be retrained, reset, replaced, duplicated, or rolled back without any loss borne by the system as such. Its identity does not depend on epistemic success. Ontological continuity is guaranteed by design rather than maintained through inference and action.

This remains true even for systems operating autonomously over extended temporal horizons. As §2 emphasises, temporal extension alone is insufficient. What matters is whether the system's own persistence as an agent depends on epistemic arbitration. In current artificial systems, it does not.

5.3 Arbitration without unified epistemic perspective

Artificial systems routinely resolve conflicts between options, evaluate trade-offs, and select actions under uncertainty. However, this does not satisfy the third continuity condition: unified arbitration at the level of the system as a whole under identity-threatening uncertainty.

In artificial systems, arbitration is modular, instrumental, and decomposable. Conflicts are resolved by optimisation routines, loss functions, or hierarchical control architectures that can be modified, replaced, or bypassed without threatening the system's persistence. Fragmentation is inexpensive. Subsystems can operate semi-independently without generating incoherence at the level of an enduring agent, because no such agent must be preserved.

By contrast, in epistemic agents fragmentation is costly. Competing subsystems cannot simply optimise independently without risking loss of diachronic unity. Consciousness supplies a unified epistemic perspective precisely because such unity is required where continuity is at stake. Artificial systems do not require this form of unity, because their operation does not depend on preserving themselves as the same epistemic subject across time.

5.4 Why scaling intelligence does not bridge the gap

It is sometimes suggested that sufficiently advanced artificial systems might acquire epistemic agency simply by scaling existing capacities. On the present account, this expectation conflates performance with structural role. Epistemic agency is not a quantitative extension of intelligence, but a qualitative shift in the relation between knowledge and continuity.

Scaling intelligence improves task-relative performance. It does not, by itself, alter the system's relation to its own persistence. Unless an artificial system must rely on its own epistemic activity to maintain itself as the same agent across time—thereby satisfying all three continuity conditions identified in §2—no increase in computational power, representational richness, or learning capacity will generate epistemic agency. The relevant distinction is functional, not incremental.

This diagnosis explains why debates about artificial consciousness often talk past one another. Performance-based arguments and phenomenological intuitions address different explanatory targets. The continuity-based account instead asks whether the system faces the problem that consciousness and epistemic agency evolved to solve. At present, artificial systems do not.

5.5 Conditions for artificial epistemic agency in principle

Nothing in this account rules out artificial epistemic agency in principle. However, the conditions under which it could arise are structural rather than merely technological. An artificial system would need to satisfy all three continuity conditions specified in §2: it would need to persist as the same agent across time without external stipulation; its continued existence would need to depend on its own epistemic arbitration; and epistemic failure would need to pose a genuine threat to its identity or viability as that agent.

This requirement should not be conflated with closing so-called control gaps. Even a system that is fully transparent, predictable, and controllable from the outside may still lack epistemic agency if epistemic failure does not accrue to the system itself. Control gaps and continuity gaps therefore come apart: closing the former does not automatically generate the latter (Veluwenkamp & Hindriks 2024).

Such a system would also require unified policy-selection under epistemic uncertainty that cannot be externally imposed, trivially reset, or modularly decomposed. Epistemic failure would need to be its failure, with consequences borne by the system itself rather than displaced onto external stakeholders. Only under these conditions would epistemic ownership—and thus epistemic agency—become functionally necessary.

Whether systems meeting these conditions are desirable, ethically acceptable, or appropriate objects of governance is a distinct question, deliberately bracketed here. The present claim is narrower and architectural: current artificial systems do not meet these conditions, and therefore do not qualify as epistemic agents under the ontological continuity condition.

5.6 Interim conclusion

Artificial systems can possess knowledge without being epistemic agents. They store information, update representations, and guide action, but they do so without epistemic ownership, ontological continuity risk, or unified arbitration under identity threat. Each of these absences corresponds directly to failure to satisfy one of the continuity conditions specified in §2.

As a result, consciousness is unnecessary for their operation. The distinction between knowledge and epistemic agency thus explains why artificial intelligence can advance rapidly without approaching the threshold at which consciousness becomes functionally required. The final section addresses potential objections to this conclusion and clarifies the limits of the continuity-based account.

6. Objections and Clarifications

The account of epistemic agency advanced in this article is deliberately restrictive. It draws a principled boundary between systems that merely instantiate knowledge and systems that qualify as epistemic agents in virtue of maintaining ontological continuity under uncertainty. Restrictive accounts invite predictable objections. This section addresses the most likely concerns and clarifies the scope and limits of the proposal. In doing so, it emphasises that the account is not intended to supplant existing theories of consciousness or agency, but to assess their adequacy with respect to a specific explanatory task: supplying a principled boundary for epistemic agency. As argued in §5, this task-relative focus establishes architectural openness rather than technological pessimism.

6.1 Is this covert anthropocentrism?

A common worry is that the continuity-based account merely redescribes human cognitive architecture and elevates it to a general standard. If epistemic agency is tied to consciousness, unified perspective, and long-horizon self-maintenance, is the result simply an anthropocentric benchmark in functional dress?

The answer is no. The criterion is not species-relative, but problem-relative. It does not privilege human capacities as such, but identifies a class of functional pressures under which epistemic agency becomes necessary. Any system—biological or artificial—that faced the same pressures would, on this account, require consciousness and qualify as an epistemic agent. Conversely, many human cognitive processes fail to meet the criterion and are therefore not agentive in the relevant sense. Habitual action, automatic inference, and procedural skill operate without epistemic ownership even in humans.

Anthropocentrism arises when human capacities are treated as normative endpoints rather than as contingent responses to particular problems. The continuity condition avoids this by grounding epistemic agency in the demands of persistence under uncertainty, not in human-like cognition, language, or social practice.

6.2 What about collective and institutional agents?

Another objection concerns collective entities such as corporations, governments, or scientific institutions. These entities appear to possess beliefs, revise them, and act on the basis of shared knowledge. If epistemic agency requires continuity and answerability across time, why do such collectives not qualify?

The continuity condition offers a clear answer. While institutions exhibit diachronic persistence, their continuity is derivative rather than internally maintained. Institutional identity is sustained through legal frameworks, social conventions, and the coordinated actions of individual agents. Epistemic error at the institutional level does not, by itself, threaten the institution's existence in the way that error threatens a biological agent's viability. Institutions can be restructured, dissolved, or reconstituted without any system-level process of self-preserving epistemic arbitration occurring within the institution itself.

This does not deny that institutions can be held epistemically responsible in derivative or normative senses. It denies only that they are epistemic agents in the same sense as conscious organisms. Their apparent agency is parasitic on the epistemic agency of their members, not emergent from an internally unified epistemic perspective.

6.3 What about unconscious belief revision?

A further concern is that much belief revision occurs without conscious awareness. Empirical work in psychology and neuroscience shows that perceptual updating, learning, and even complex inferential processes can proceed unconsciously. Does this undermine the claim that consciousness is required for epistemic agency?

It does not. The present account does not claim that all epistemic processing must be conscious. Rather, it claims that epistemic agency as a whole—the capacity to own, revise, and act upon knowledge as the same agent across time—requires a conscious integrative perspective. Unconscious processes can support epistemic agency without constituting it. They operate as subpersonal mechanisms within a system whose overall epistemic activity is unified and answerable.

This distinction mirrors familiar distinctions in action theory. Many motor and perceptual processes are unconscious, yet agency is not thereby eliminated. What matters is not the consciousness of every process, but the presence of a conscious perspective that integrates those processes into coherent belief and action over time.

6.4 Is epistemic agency being conflated with moral responsibility?

Another potential misreading is that epistemic agency is being treated as a moral or normative category. Because epistemic agency involves answerability, error, and responsibility, it may appear that the account smuggles in ethical claims about blame or obligation.

This is a mistake. The responsibility at issue is epistemic, not moral. To be epistemically responsible is to be a locus of belief revision under risk, not to be a bearer of moral duties. The account does not entail that epistemic agents deserve praise or blame, nor that non-agents lack moral standing. It simply identifies the conditions under which epistemic practices such as justification, critique, and revision make sense as activities of a persisting agent.

Separating epistemic agency from moral responsibility is essential to avoiding overreach. The continuity condition is a descriptive constraint on when epistemic agency arises, not a normative claim about how agents ought to be treated.

6.5 Relation to competing theories of consciousness and agency

The discussion of large language models in §4.7 should be read as a diagnostic illustration of the continuity condition rather than as a special or motivating case; the same constraint applies uniformly across biological and artificial systems.

The continuity-based account should not be read as refuting existing theories of consciousness or agency. Rather, it evaluates their adequacy with respect to a specific explanatory task: supplying a principled stopping point for epistemic agency.

Panpsychist accounts, for example, aim to address the metaphysical status of consciousness by treating it as fundamental or ubiquitous. While such accounts may succeed at other explanatory tasks, they do not, by themselves, supply a criterion for why epistemic agency would be unnecessary in systems whose continuity is structurally guaranteed and necessary in systems facing long-horizon epistemic risk. In this respect, panpsychism remains incomplete for the present boundary-setting task, even if not for others.

Higher-order theories succeed in distinguishing conscious from unconscious mental states within mature cognitive systems, but they do so by specifying mechanisms rather than necessity conditions. As a result, they draw cognitively specific boundaries without explaining why metarepresentation should be required for epistemic agency in the first place. They describe how agency may be implemented in some systems, but do not, by themselves, explain why it becomes necessary at all.

Integrated Information Theory and global workspace theories offer particularly instructive contrasts. Both emphasise large-scale informational integration and broadcasting as central to conscious processing, and both are supported by substantial empirical and theoretical work in cognitive neuroscience (Koch et al. 2016; Mashour et al. 2020). These frameworks are valuable for explaining differences between conscious and unconscious processing within organisms, and for identifying neural correlates and mechanisms of conscious access.

However, they do not, by themselves, address the further question at issue here: why such integration should become necessary for a system as a persisting epistemic agent. Informational integration may be sufficient for certain forms of conscious processing or reportability, but without an explicit continuity-based constraint it does not explain why epistemic ownership and answerability arise in some systems and not others.

The present account is therefore complementary rather than competitive. It does not deny the relevance of integration, broadcasting, or workspace dynamics. It claims only that these mechanisms require an additional constraint—ontological continuity under epistemic risk—to explain why epistemic agency appears where it does and not elsewhere.

6.6 Does the account trivialise knowledge?

Finally, one might worry that distinguishing knowledge from epistemic agency trivialises knowledge by allowing it to proliferate without constraint. If databases, algorithms, and simple biological systems can all possess knowledge, does the concept lose its significance?

This worry confuses scope with depth. Allowing knowledge to exist without consciousness does not flatten epistemic distinctions; it sharpens them. It allows us to explain why some forms of knowledge are shallow, derivative, or externally maintained, while others are deep, owned, and consequential for the agent's future. Epistemic agency marks this depth distinction. It identifies when knowledge is not merely present, but at stake for a persisting system.

In this respect, the account preserves what matters about epistemic practice while resisting unnecessary inflation. Knowledge can be widespread; epistemic agency is rare. Consciousness, on this view, marks the threshold between the two.

6.7 Summary clarification

The ontological continuity constraint advanced here is orthogonal to metaphysical theories of consciousness: it neither presupposes nor adjudicates between accounts of what consciousness is, but specifies a functional condition under which consciousness becomes necessary for epistemic agency.

The objections considered here do not undermine the continuity-based account of epistemic agency. They instead clarify its scope and ambition. The account is not anthropocentric, not moralised, and not threatened by unconscious processing. Nor does it deny the value of competing theories for other explanatory purposes. Its contribution is narrower but precise: it identifies the conditions under which epistemic ownership, revision, and answerability become necessary at all.

With these clarifications in place, the concluding section draws together the article's contributions and situates them within broader debates in epistemology, philosophy of mind, and artificial intelligence.

7. Conclusion: Consciousness as a Threshold for Epistemic Responsibility

This article has argued for a principled distinction between knowledge and epistemic agency. Knowledge, understood as information that can guide behaviour, can be instantiated in a wide range of systems, including non-conscious biological mechanisms and artificial systems. Epistemic agency, by contrast, is the capacity to hold, revise, and act upon knowledge as the same agent across time. The central claim has been that epistemic agency presupposes consciousness, not because consciousness confers intelligence or rationality, but because it enables the maintenance of ontological continuity under epistemic risk.

By grounding epistemic agency in the ontological continuity condition, the article reframes consciousness as a functional threshold rather than a metaphysical essence. Consciousness marks the point at which knowledge must be owned, defended, and revised by a persisting agent whose future depends on the success of its epistemic arbitration. This explains why intelligence, learning, and planning can exist without consciousness, while epistemic responsibility cannot. Consciousness, on this view, is not a general enhancer of cognition, but a structural requirement for epistemic answerability.

This distinction has implications across several domains. In epistemology, it clarifies why epistemic practices—such as justification, critique, revision, and accountability—presuppose agents capable of maintaining identity across time. These practices make sense only where error can be attributed to a persisting subject and corrected in light of future consequences. In this context, accounts of epistemic responsibility are invoked diagnostically rather than normatively, in order to illuminate the structural presuppositions of epistemic practice rather than to advance claims about epistemic injustice, blame, or moral standing (Fricker 2007, pp. 130–135). In philosophy of mind, the account complements existing theories by supplying a boundary condition: consciousness becomes functionally necessary not wherever information is integrated, but where epistemic failure threatens continuity.

In debates about artificial intelligence, the framework supplies a missing structural explanation for concerns that are already widely recognised but insufficiently grounded. Contemporary discussions of responsibility gaps, accountability diffusion, and moral overload in AI systems frequently presuppose agents without explaining why agency fails to arise in artificial systems despite their growing epistemic competence (Gogoshin 2025). The continuity-based account advanced here does not resolve those debates, but it clarifies their underlying architecture: responsibility does not attach merely because systems influence outcomes, but because epistemic failure must be borne by the system itself across time. Where such self-risk is absent, responsibility gaps are not anomalies to be patched, but predictable structural features.

It is important to emphasise what does not follow from this account. The argument does not entail moral, legal, or political conclusions about how epistemic agents ought to be treated, nor does it deny moral standing or ethical value to non-conscious systems. The responsibility at issue is epistemic rather than ethical. Likewise, the article does not aim to resolve the hard problem of consciousness or to adjudicate between competing metaphysical theories of mind. Its contribution is deliberately limited to identifying a necessary condition for epistemic agency—a constraint that is orthogonal to disputes about the ultimate nature of conscious experience.

By distinguishing knowledge from epistemic agency and locating consciousness at the threshold between them, the article offers a way of preserving the significance of epistemic responsibility without inflating the concept of consciousness. Consciousness emerges neither as a mysterious surplus nor as a mere by-product of intelligence, but as a condition that arises where—given the continuity problem—it becomes functionally necessary on the present account for a system to treat its epistemic states as its own across time. In this respect, consciousness marks the boundary not between simple and complex cognition, but between systems that merely process information and those that must answer for it diachronically (Clark 2016, pp. 267–270).

Together with the companion article developing the ontological continuity condition for consciousness, the present argument forms a two-stage constraint–application programme: first identifying when consciousness becomes necessary in principle, and then showing how that necessity structures epistemic agency. Future work may extend this programme by operationalising continuity-related constraints empirically, examining their relevance for AI governance and system design, and exploring how epistemic responsibility is distributed, delegated, or obscured within institutional and socio-technical settings.

References

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227–247.

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190217013.001.0001

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7–19. https://doi.org/10.1093/analys/58.1.7

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology*, 18(4), 299–309. https://doi.org/10.1007/s10676-016-9403-3

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

Dretske, F. (1981). *Knowledge and the flow of information*. MIT Press.

Dreyfus, H. L. (2002). Intelligence without representation: Merleau-Ponty's critique of mental representation. *Phenomenology and the Cognitive Sciences*, 1(4), 367–383. https://doi.org/10.1023/A:1021351606209

Fitts, P. M., & Posner, M. I. (1967). *Human performance*. Brooks/Cole.

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198237907.001.0001

Gallagher, S. (2000). Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences*, 4(1), 14–21. https://doi.org/10.1016/S1364-6613(99)01417-5

Gendolla, G. H. E. (2015). Implicit affect primes effort: Basic processes, moderators, and boundary conditions. *Social and Personality Psychology Compass*, 9(11), 606–619. https://doi.org/10.1111/spc3.12208

Gogoshin, D. L. (2025). A way forward for responsibility in the age of AI. *Inquiry*, 68(4), 1164–1197. https://doi.org/10.1080/0020174X.2024.2312455

Janeway, C. A., Travers, P., Walport, M., & Shlomchik, M. J. (2001). *Immunobiology: The immune system in health and disease* (5th ed.). Garland Science.

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience* 17, 307–321. https://doi.org/10.1038/nrn.2016.22

Mashour, G. A., Roelfsema, P., Changeux, J. P., & Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron*, *105*(5), 776–798. https://doi.org/10.1016/j.neuron.2020.01.026

Medzhitov, R., & Janeway, C. A., Jr (2002). Decoding the patterns of self and nonself by the innate immune system. *Science*, *296*(5566), 298–300. https://doi.org/10.1126/science.1068883

Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. MIT Press.

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

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, 10(5), 233–238. https://doi.org/10.1016/j.tics.2006.03.006

Reber, A. S. (1993). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious*. Oxford University Press.

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems. *Frontiers in Robotics and AI*, *5*, 15. https://doi.org/10.3389/frobt.2018.00015

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–457. https://doi.org/10.1017/S0140525X00005756

Shepherd, J. (Ed.). (2022). *The Routledge handbook of philosophy of agency*. Routledge.

Veluwenkamp, H., & Hindriks, F. (2024). Artificial agents: Responsibility and control gaps. *Inquiry*. https://doi.org/10.1080/0020174X.2024.2410995

Zahavi, D. (2005). *Subjectivity and selfhood: Investigating the first-person perspective*. MIT Press. https://doi.org/10.7551/mitpress/6541.001.0001