https://doi.org/10.65166/w8zvx652

International Journal of Behavioral and Social Analytics
Volume I, Issue 2, May 2026
eISSN: 3116-4358
https://journal.ijhba.com/index.php/ijbesa
https://slsipress.com

From Syllogism to Statistical Inference: Machine Learning as Epistemological Rupture and the Future of Human Cognitive Evolution

Ramon George Atento, PhD and Leah Quinto, PhD

Abstract

For nearly three millennia, Western epistemology has framed human knowledge as a rational, intentional, and structurally ordered process — from Aristotle's syllogistic logic through Kant's constitutive categories of understanding (Bond, 2021; Russon, 2016). This paper argues that the emergence of machine learning constitutes a fundamental epistemological rupture with that tradition rather than an extension of it. Where classical knowing moves from principles to conclusions through intentional, phenomenologically grounded processes, machine learning moves from data to emergent pattern through procedures that are non-intentional, phenomenologically empty, and in significant measure opaque (Barelli et al., 2024; Lykhatskyi, 2025). This structural inversion is not a deficiency of machine learning — it is a philosophically significant divergence that the classical tradition lacks the conceptual vocabulary to evaluate and that demands new integrative frameworks to understand. Drawing on extended mind theory, cognitive offloading research, and evolutionary epistemology, this paper develops such a framework, arguing that machine learning represents the most consequential instance of cognitive extension in human intellectual history — and the first to introduce systemic opacity into the extended human cognitive system. This opacity extension generates specific behavioral vulnerabilities, including automation bias, skill atrophy, and epistemic dependence that are structural properties of the human-AI cognitive relationship rather than correctable traits of individual users (Natali et al., 2025; Zerilli et al., 2019). The paper further argues, through an evolutionary-epistemology lens, that the current moment of AI integration constitutes a transition point in the adaptive history of human cognitive ecology whose long-term consequences for epistemic identity, collective knowledge-making, and human cognitive development remain critically undertheorized. This paper employs a philosophical-psychological conceptual synthesis methodology, integrating philosophy of mind, cognitive science, behavioral research, and educational literature. The paper concludes with recommendations for research, educational design, and institutional governance, calibrated to the structural properties of AI-integrated cognition.

Keywords: *epistemological inversion; machine learning and human cognition; extended mind theory; evolutionary epistemology; cognitive offloading; epistemic identity; AI-integrated cognition*

Article History: Received March 12, 2026 | Revised April 3, 2026 | Accepted May 2, 2026

1. Introduction

Few questions in the history of human thought have proven as enduring — or as consequential — as the question of how human beings come to know: From Aristotle's systematic account of syllogistic reasoning to Kant's mapping of the mind's constitutive categories, Western epistemology has proceeded on a foundational assumption: that knowledge is rational, intentional, and structurally ordered. To know something, in the classical sense, is not merely

to have encountered it — it is to have grasped it through the disciplined exercise of reason, moving from premises to conclusions, from principles to particulars, from evidence to judgment. This picture of the knowing mind — deliberate, hierarchical, and phenomenologically grounded in a subject who understands — has shaped not only philosophy but psychology, education, and the social organization of expertise for nearly three millennia (Bond, 2021; Russon, 2016).

In roughly a decade, that picture has been placed under extraordinary pressure. The emergence of machine learning, and more specifically the rapid development of large-scale deep learning systems, has produced artifacts that perform — with striking and in some domains superhuman competence — many of the cognitive functions that classical epistemology reserved for rational minds (Aithal, 2023). These systems classify, infer, generate, translate, diagnose, and predict (Oermann et al., 2025). They do so not by following explicit rules derived from principles, as classical artificial intelligence once attempted, but by extracting statistical regularities from vast quantities of data through processes that are, in many cases, opaque even to their designers (Peters, 2022). The knowledge these systems produce — if it can be called knowledge at all — does not arise from intention, understanding, or phenomenological engagement with the world. It arises from pattern, probability, and scale.

This is not merely a technical development. It is an epistemological event. The behavioral and social implications of this event are already visible and are intensifying ("Embracing the Ubiquity of Machines," 2024). Human beings are increasingly delegating cognitive functions to machine learning systems — not only computational tasks, but interpretive ones (Dergaa et al., 2024). Physicians consult AI diagnostic tools (Krishnan et al., 2023). Judges and parole boards work alongside algorithmic risk assessments (Ramos et al., 2019). Students compose, research, and reason with AI assistance (Black & Tomlinson, 2025). Researchers use machine-generated literature syntheses as starting points for inquiry (Block & Kuckertz, 2024). In each of these contexts, the classical boundary between human knowing and machine processing is not simply being tested — it is being actively renegotiated in practice, often faster than theory can follow.

Psychology and cognitive science have begun to document the behavioral consequences of this renegotiation (Williams & Lim, 2024). Studies of cognitive offloading suggest that human beings readily and rapidly adapt their reasoning processes when reliable external cognitive tools are available, redistributing mental effort in ways that alter both the content and the process of thinking (Dergaa et al., 2024). The extended mind thesis, developed by Andy Clark and David Chalmers, offers a theoretical framework for understanding this adaptation not as cognitive laziness or epistemic surrender, but as a continuation of a pattern that has defined human intelligence throughout its evolutionary history — the systematic extension of cognitive capacity beyond the biological boundaries of the individual brain through tools, symbols, institutions, and now, increasingly, machines (Dergaa et al., 2024).

This paper takes that framework seriously and pushes it further. It proposes that the emergence of machine learning is not best understood as a rupture with human cognition, nor as a seamless extension of it, but as a philosophically complex inversion that demands new conceptual vocabulary. Where classical epistemology moves from principles to conclusions, ML moves from data to emergent pattern (Andrews, 2025; Barelli et al., 2024). Where human knowing — in its classical description — is intentional and phenomenologically grounded, ML is non-intentional and phenomenologically empty (Orbik, 2024; Sims, 2021). Yet where human reasoning is slow, resource-intensive, and subject to well-documented cognitive biases, ML at scale is fast, consistent, and capable of detecting patterns that exceed unaided human perceptual capacity (Diržytė, 2025; Page & Kallapur, 2025). These are not symmetric differences. They raise the question of whether what human beings are building in machine learning systems constitutes a genuinely new mode of epistemic operation — and if so, what that means for the long-term trajectory of human cognition as a species-level adaptive capacity.

The paper pursues this question through an integrative philosophical-psychological conceptual synthesis, drawing on the philosophy of mind, cognitive science, evolutionary epistemology, and behavioral research on human-AI interaction. It does not claim that machine learning systems know, understand, or experience in any phenomenological sense. It does claim that their existence and proliferation constitute a significant event in the history of human knowing — one that alters the behavioral environment within which human cognition operates, and that carries implications for education, epistemic trust, identity, and the social organization of intelligence that the scholarly literature has not yet fully addressed.

The specific objectives of this paper are as follows: First, it examines the classical epistemological tradition to establish the framework against which ML must be assessed, identifying its core assumptions about rationality, intentionality, and the structure of knowing. Second, it analyzes machine learning as an epistemic mode, distinguishing its operations from classical cognition and interrogating the meaning of the black box problem as a philosophical

rather than merely technical challenge. Third, it applies Extended Mind Theory and the cognitive offloading literature to situate ML within the longer history of human cognitive extension, assessing the degree to which AI represents a qualitative departure from prior forms of cognitive tool use. Fourth, it draws on evolutionary epistemology to consider whether ML can be read as a new phase in the adaptive development of human knowledge-making capacity. Fifth, it examines the behavioral and educational implications of AI-integrated cognition, grounding the paper's more philosophical arguments in observable consequences for human learning and epistemic practice. Finally, it synthesizes these threads into a coherent conceptual framework and proposes directions for future empirical and qualitative inquiry.

The stakes of this inquiry extend beyond academic philosophy. As machine learning systems become embedded in the infrastructures of knowledge production, education, professional judgment, and social decision-making, the question of how they relate to human cognition is not merely interesting — it is urgent (Giovanni, 2025; Webb et al., 2020). The framework this paper develops contributes to that urgency: an attempt to think carefully, at the intersection of philosophy, psychology, and behavioral science, about what kind of knowing is now possible, what kind remains distinctively human, and what the relationship between the two will mean for how human beings understand themselves as thinking beings in the decades and centuries ahead.

2. Review of Related Literature

2.1 Classical Epistemology and the Architecture of Human Knowing

The Western epistemological tradition has, across its long development, maintained a remarkably consistent set of assumptions about the nature of genuine knowledge (Botha et al., 2021; Wilburn, 2020). To know something, in the sense that philosophers from Aristotle onward have found philosophically significant, is not merely to have reliable access to correct information. Instead, it involves standing in a particular cognitive relationship to that information — one characterized by rational justification, intentional engagement, and the capacity to account for why something is the case rather than merely that it is (Agarwal, 2017; Toffoli, 2024). This architecture of knowing, built over centuries and refined through successive philosophical debates, constitutes the baseline against which the epistemological status of machine learning must ultimately be assessed.

Aristotle's contribution to this architecture was foundational in ways that continue to reverberate. His account of syllogistic reasoning established the formal structure of valid inference, and his distinction between episteme — scientific knowledge grounded in necessary causes — and doxa — mere opinion — drew a consequential boundary between knowing and guessing that Western philosophy has largely preserved. For Aristotle, genuine knowledge required not only true belief but demonstration: the capacity to show why something could not be otherwise, grounded in first principles apprehended through nous, a form of rational intuition that could not itself be derived from syllogistic procedure (Aristotle, 1994). This combination of formal inferential structure and intuitive rational grasp established a model of knowing that was simultaneously logical and phenomenologically grounded in a subject capable of understanding.

Descartes radicalized this picture by making the certainty of the knowing subject the irreducible foundation of all knowledge. In severing reliable knowledge from sensory experience and grounding it in the pure operations of reason, Descartes both intensified the rationalist strand of the tradition and made the first-person perspective of the knowing mind constitutively central to epistemology in a way that would prove enormously influential (Cottingham, 2016). The cogito is not merely a logical argument; it is a claim about the indispensability of conscious rational subjectivity to the very possibility of knowledge. This move has significant implications for any assessment of ML, which produces outputs through processes entirely devoid of such subjectivity.

Kant's intervention complicated and deepened the tradition by arguing that knowledge is neither purely rational nor purely empirical but arises from the active synthesis of sensory intuition under the constitutive categories of the understanding (Yamamoto, 2017). For Kant, the mind does not passively receive knowledge from the world; it actively structures experience through innate cognitive forms that make ordered experience possible (Kant, 1998). This constructivist turn, while it opened epistemology to a richer engagement with experience, preserved the centrality of rational structure and the necessity of a transcendental subject whose cognitive activity makes knowledge possible. The knower, for Kant, is not incidental to the known — the knowing mind constitutes the epistemic order within which objects of knowledge can appear.

Contemporary philosophy of mind has both extended and challenged this tradition in ways relevant to the present paper. Phenomenological thinkers, particularly Husserl and Heidegger, deepened the tradition's insistence on intentionality — the directedness of conscious mental states toward objects — as a non-negotiable feature of genuine cognition (Heidegger, 1927/1962; Husserl, 1900/1970). Heidegger's account of understanding as embedded, practical, and world-involving provided the philosophical foundation for Hubert Dreyfus's influential critique of classical AI, which argued that rule-based computational systems could never replicate the embodied, contextual intelligence that characterizes human knowing because they lacked the existential situatedness that makes genuine understanding possible (Dreyfus, 1978).

Yet a significant line of contemporary philosophy has pushed back against what it regards as an overly restrictive and mentalist conception of knowledge. Externalist epistemologists, including Alvin Goldman and Fred Dretske, have argued that knowledge does not require the kind of rational transparency the classical tradition demands — that a belief can constitute knowledge if it is produced by a reliable cognitive process, regardless of whether the knower can articulate the justification (Hatfield & Goldman, 1989). This reliabilist turn is philosophically significant for this paper because it opens a conceptual space in which the outputs of ML systems might, under certain conditions, satisfy epistemological criteria without requiring intentionality, phenomenological engagement, or rational transparency. The tension between internalist and externalist accounts of knowledge in the contemporary literature is thus not merely an academic dispute — it is directly relevant to how one evaluates what ML systems produce and whether that production can meaningfully be called knowing.

What emerges from this survey is a tradition that is internally more contested than its classical formulations suggest, but that has consistently, across its major variations, treated the rational, intentional, and phenomenologically grounded subject as either constitutive of or at least necessary for genuine knowledge. It is precisely this consistency that machine learning places under pressure.

2.2 Machine Learning as a Mode of Knowledge Acquisition: Convergence and Divergence

The relationship between machine learning and the classical epistemological tradition is one of simultaneous superficial resemblance and deep structural divergence. On the surface, ML systems appear to do something epistemically familiar: they take in information, process it, and produce outputs that function as representations of patterns in the world. They improve with experience, in the sense that their outputs become more accurate as they are exposed to more data. In some domains — medical imaging, protein folding, natural language processing — their representational accuracy now rivals or exceeds that of human experts (Alzubaidi et al., 2021; Pellegrain, 2023). These surface features invite the interpretation that ML is a particularly powerful form of the kind of inductive learning that empiricist epistemologists have long regarded as central to human knowledge acquisition.

The divergence, however, is fundamental and becomes apparent upon closer inspection. Classical rule-based AI, which dominated the field from its inception through the 1980s, was explicitly designed based on an epistemological model continuous with the rationalist tradition (Larsen, 2022; Maclure, 2021). Expert systems encoded human knowledge in the form of explicit rules and logical inference procedures, and their operations were in principle fully transparent — every output could be traced through a chain of explicitly represented inferential steps (Russell & Norvig, 1995). This made classical AI epistemologically tractable, even if ultimately limited. Dreyfus's critique applied directly to this paradigm: systems that operate through explicit symbol manipulation and rule-following cannot replicate the contextual, embodied, and tacit dimensions of human expertise, which depend on forms of understanding that resist formalization.

Modern deep learning operates on an entirely different basis and is, in important respects, immune to Dreyfus's specific objections while raising new and more profound epistemological questions of its own. Deep neural networks do not operate through explicit rules. They learn through exposure to vast quantities of training data, adjusting billions of numerical parameters through optimization procedures that minimize prediction error across the training distribution. The representations that emerge from this process are distributed across the network in ways that are not interpretable as discrete symbolic rules or logical propositions. The system learns to recognize cats, diagnose tumors, or generate coherent text not by applying explicit criteria but by developing internal representations whose functional properties are not accessible to straightforward inspection (LeCun et al., 2015).

This opacity is the source of what has come to be called the black box problem, and the present paper argues, following a line of philosophical analysis developed by scholars including Burrell and Pasquale, that it is not merely a technical inconvenience but a philosophically significant epistemic challenge (Burrell, 2016; Pasquale, 2015). The

classical epistemological tradition, in virtually all its variants, has treated the intelligibility of the inferential process as integral to the epistemic status of the output. For Aristotle, demonstration requires that the inferential chain be traceable to first principles. For Kant, knowledge requires that the synthesis of intuition under concepts be a process that the transcendental subject performs. For internalist epistemologists, justified belief requires that the justification be accessible to the believer. A system whose outputs cannot be traced to interpretable inferential processes raises the question not merely of whether we can trust those outputs, but of what kind of epistemic object they are.

John Searle's Chinese Room argument, while developed in a different context and directed at a different target, illuminates this problem from a complementary angle. Searle's argument that syntactic symbol manipulation, however sophisticated, does not constitute semantic understanding points to a deeper issue: that the production of correct or useful outputs does not entail understanding of the domain from which those outputs are drawn (Searle, 1980). Applied to modern ML, the argument suggests that a system that produces accurate medical diagnoses, compelling philosophical arguments, or correct mathematical proofs need not thereby understand medicine, philosophy, or mathematics in any sense consistent with what the classical tradition has meant by understanding. The outputs and the understanding have been decoupled in a way that the tradition did not anticipate and for which it lacks a ready conceptual vocabulary. In adjacent health analytics work, Atento, Quinto, Espelita, and San Juan (2025) make a parallel caution that computational processing of human narratives requires interpretability, cultural mediation, and ethical governance if meaning-rich experience is not to be flattened into decontextualized data signals.

Recent debates in AI research and philosophy of AI have complicated this picture in productive ways. Scholars including Gary Marcus and Ernest Davis have argued that current ML systems, despite their impressive performance, fail to achieve genuine generalization — they interpolate within their training distributions rather than reasoning from principles, and they fail in systematic ways when confronted with situations that fall outside those distributions (Marcus & Davis, 2019). This failure mode is itself epistemologically revealing: it suggests that what ML produces is better described as compressed pattern representation than the kind of principled, generalizable knowledge that the classical tradition has valued. Others have argued that the representational structures developed by some large-scale models may constitute a form of world-modeling with genuine epistemic significance that current philosophical vocabulary is inadequate to describe (Mai et al., 2024). However, Yann LeCun has also expressed skepticism about current large language models' ability to achieve human-level intelligence due to their lack of true reasoning and understanding of the physical world (Gurevich & Blass, 2024).

What the literature in this area establishes, across its disagreements, is that machine learning constitutes a genuinely novel epistemic phenomenon that neither fits comfortably within the classical tradition nor can be dismissed as epistemologically trivial. Its relationship to human knowing is one of structural divergence rather than mere quantitative difference, and the implications of that divergence for how human beings understand knowledge, expertise, and intelligence are only beginning to be theorized.

2.3 Extended Mind Theory and Cognitive Offloading

The philosophical framework most directly relevant to understanding the relationship between human cognition and machine learning as a behavioral and psychological phenomenon is the extended mind thesis, developed by Andy Clark and David Chalmers in their landmark 1998 paper and elaborated across subsequent decades of philosophical and cognitive scientific work. The core claim of the extended mind thesis is deceptively simple but philosophically far-reaching: the boundary of the cognitive system is not fixed at the boundary of the biological organism. When an external resource — a notebook, a calculator, a smartphone, or potentially an AI system — plays a sufficiently reliable, accessible, and automatically endorsed role in a cognitive process, it constitutes a genuine component of that process rather than merely an input to it (Clark & Chalmers, 1998).

The thought experiment Clark and Chalmers use to motivate this claim — the case of Inga, who remembers the address of a museum by consulting her biological memory, compared with Otto, who has Alzheimer's and keeps the same information in a notebook he consults automatically — is designed to show that the functional equivalence of the two processes is sufficient to warrant treating the notebook as part of Otto's cognitive system in the same sense that Inga's biological memory is part of hers (Clark & Chalmers, 1998). The philosophical significance of this argument for the present paper is considerable: if the criterion for cognitive extension is functional integration rather than biological implementation, then the systematic use of ML tools in knowledge work raises genuine questions about where the human cognitive system ends and the machine system begins.

Clark's subsequent work, particularly in Being There and Natural-Born Cyborgs, developed this framework into a broader account of human beings as constitutively tool-using, tool-incorporating cognitive agents (Clark, 2003). On this account, the history of human intelligence is inseparable from the history of cognitive scaffolding — the progressive elaboration of external structures that extend, amplify, and reorganize human cognitive capacity (Clark, 2001). Writing transformed not only what human beings could remember but also how they could think. Mathematical notation made possible forms of reasoning that would be inaccessible to unaided biological cognition. Digital computation extended human processing capacity across domains from engineering to genomics (Heersmink & Knight, 2018). Each of these developments altered the cognitive environment within which human intelligence operates, and each provoked versions of the anxieties now associated with AI — concerns about cognitive dependence, the erosion of intrinsic capacities, and the proper boundaries of authentic human knowing (Clark, 2025).

The cognitive offloading literature provides empirical grounding for the extended mind framework by documenting the behavioral mechanisms through which this extension occurs. Risko and Gilbert, in a widely cited review, characterize cognitive offloading as the use of action to alter the informational state of the environment to reduce the cognitive demands of a task (Risko & Gilbert, 2016). Research in this area has consistently demonstrated that humans are flexible and rapid cognitive offloaders: they redistribute cognitive effort to external resources whenever those resources are perceived as reliable and accessible, and they do so in ways that alter not only the efficiency but the qualitative character of their cognitive engagement with the task (Grinschgl et al., 2020; Grinschgl & Neubauer, 2022). Studies of GPS navigation, search engine use, and calculator reliance have all documented that offloading changes not only performance outcomes but also the development of intrinsic capacities — in some cases preserving them by freeing cognitive resources for higher-order tasks, in others eroding them through disuse (Cassinadri & Fasoli, 2023; George et al., 2024).

The application of this literature to AI integration is a growing area of research, though it remains underdeveloped relative to the scale of the phenomenon. Preliminary findings suggest that the use of AI writing and reasoning assistants alters how users approach cognitive tasks, including reductions in the perceived need for independent evidence evaluation, changes in the attribution of epistemic authority, and shifts in the experience of cognitive effort. These behavioral changes are not inherently problematic — they may represent rational adaptive responses to a changed cognitive environment — but they raise important questions about the cumulative effects of deep AI integration on human cognitive development, epistemic autonomy, and the social distribution of knowledge competence.

A significant philosophical debate within the extended mind literature concerns whether AI systems satisfy the conditions Clark and Chalmers specify for genuine cognitive extension, or whether they constitute a qualitatively different kind of relationship. Critics including Adams and Aizawa have argued that the extended mind thesis conflates the causal coupling of cognitive and non-cognitive processes, and that genuine cognition requires intrinsic intentional content that external devices cannot possess (Adams & Aizawa, 2001). On this view, even a highly integrated AI system remains a tool rather than a component of the user's cognitive system. Defenders of the extended mind position respond that this objection relies on an unsupported assumption about the biological basis of intentionality, and that the functional integration criterion is both more tractable and more theoretically defensible.

This debate has not been resolved, but it is relevant to the present paper. Whether or not ML systems satisfy the strict conditions for cognitive extension, the behavioral evidence that human cognition is being reorganized around their availability is accumulating, and the extended mind framework — even in its more modest interpretations — provides the most theoretically sophisticated account of how that reorganization should be understood and evaluated.

2.4 Evolutionary Epistemology and the Adaptive Nature of Knowledge

Evolutionary epistemology represents a departure from the classical tradition's tendency to treat knowledge as a timeless rational achievement, proposing instead that the capacity for knowledge acquisition is a biological phenomenon shaped by the same evolutionary processes that also formed all other adaptive characteristics of living organisms (Clark, 1986). This reframing has significant implications for the questions that the present paper pursues, because it opens a temporal and developmental perspective on human knowing that the purely synchronic analyses of classical epistemology cannot provide.

The foundational contributions to evolutionary epistemology come from two distinct but related traditions. The first, associated primarily with Karl Popper, applies an evolutionary logic to the growth of scientific knowledge: just as biological evolution proceeds through the generation of variation and the selective retention of adaptive traits, so

scientific knowledge grows through the proposal of bold hypotheses and the selective retention of those that survive critical testing (Ruja & Popper, 1973). For Popper, this evolutionary analogy is epistemological rather than biological — he is not claiming that scientific rationality is reducible to biology, but that the logic of knowledge growth mirrors the logic of adaptive selection. The second tradition, associated with Konrad Lorenz and subsequently developed by Gerhard Vollmer and Henry Plotkin, takes the biological grounding more literally, arguing that human cognitive structures — including the basic categories through which humans perceive and organize experience — are themselves evolutionary adaptations, shaped by selection pressures over the long history of hominid development (Plotkin, 1995; Wilson et al., 1977).

Daniel Dennett's contribution to this area, particularly in Darwin's Dangerous Idea and Consciousness Explained, extends the evolutionary framework to encompass cultural and technological developments as well as biological ones, proposing that human intelligence is inseparable from the cultural scaffolding within which it develops and operates (Dennett, 1995). For Dennett, the tools, symbols, and institutions that humans have created are not merely aids to a pre-existing biological intelligence — they are constitutive of the distinctively human form of intelligence, which has co-evolved with its cultural and technological environment in ways that make the boundary between the biological and the artifactual difficult to draw cleanly. This position has an evident affinity with Clark's Extended Mind Thesis and provides an evolutionary perspective for the cognitive offloading framework discussed in the previous section.

The implications of evolutionary epistemology for the assessment of machine learning are significant but require careful handling. The straightforward evolutionary reading — that ML represents the next phase in the adaptive extension of human cognitive capacity — risks oversimplification by treating a complex technological development as directly continuous with the biological processes that shaped human intelligence. Plotkin's account of the hierarchy of adaptive mechanisms, which distinguishes between genetic evolution, individual learning, and cultural transmission as nested adaptive systems operating on different timescales, provides a more nuanced framework: ML might be understood as a novel component of the cultural adaptive system, generating a new form of heritable cognitive tool whose effects on human epistemic behavior will unfold on timescales that are fast relative to biological evolution but slow relative to individual learning (Plotkin, 1995).

The evolutionary epistemology literature has also engaged directly with the question of whether technological extension of human cognition constitutes adaptive progress or introduces novel vulnerabilities. Stanovich's work on rationality and its limitations, while not framed in explicitly evolutionary terms, provides relevant evidence that human cognitive architecture has systematic biases and limitations that evolved in environments very different from those in which contemporary knowledge demands operate, and that the tools humans use to compensate for these limitations can either ameliorate or amplify them depending on their design and deployment (Stanovich, 2017). This finding introduces an important qualification to any straightforwardly optimistic evolutionary reading of AI integration: the adaptive value of a cognitive tool depends on the fit between the tool's characteristics and the cognitive demands of the environment in which it is used, and this fit cannot be assumed.

Recent work in cognitive science and philosophy of technology has begun to explore whether the pace and scale of AI development introduces a qualitatively new dynamic into the co-evolutionary relationship between human cognition and its technological environment. Stiegler's account of technics as a constitutive dimension of human memory and cognition — humans are, on his account, beings whose memory and cognitive capacity are always already exteriorized in technical objects — provides a philosophical vocabulary for thinking about this dynamic, though his framework requires translation into more empirically tractable terms to engage with the behavioral literature (Stiegler, 1998). What this body of work collectively suggests is that the evolutionary epistemology framework is not merely a metaphor for understanding AI development — it is a conceptually appropriate lens for a phenomenon that involves the reorganization of human cognitive ecology at a scale and pace without historical precedent.

2.5 Behavioral and Educational Implications of AI-Integrated Cognition

The philosophical and theoretical frameworks developed in the preceding sections gain their most direct behavioral and social relevance in the domain of education and knowledge practice, where the integration of AI tools is already producing observable changes in how human beings learn, reason, and relate to knowledge. This theme brings the paper's more abstract arguments into contact with the empirical literature on human behavioral adaptation to AI and grounds the paper's contribution in consequences that matter for individuals, institutions, and societies.

The educational literature on AI integration has expanded rapidly since the public availability of large language models accelerated dramatically in 2022 and 2023 (Cibu et al., 2025; Garzón et al., 2025). Early research in this area focused primarily on questions of academic integrity and the detection of AI-generated work — concerns that, while practically important, reflect a defensive posture toward AI that the more analytically sophisticated literature has begun to move beyond (Avello & Zurita, 2025; Jensen et al., 2024). More substantive contributions have addressed the effects of AI tool availability on learning processes themselves, with findings that are complex and in some respects counterintuitive (Liu et al., 2026; Rao, 2025). Mollick and Mollick, in work examining AI integration in educational contexts, found that the availability of AI assistance could both accelerate learning in some domains and create what they describe as a risk of skill atrophy in others — specifically in domains where the productive struggle of working through difficulty without assistance is itself a mechanism of learning (Mollick & Mollick, 2023). This finding resonates directly with the cognitive offloading literature's documentation of the double-edged nature of external cognitive support. In the Philippine higher education context, related work has also found that e-learning adoption may be widespread yet framed primarily as instructional continuity rather than as a deeper transformation of curriculum, pedagogy, and institutional strategy (Atento, 2025). A related survey by Rao, Tian, and Atento (2025) similarly found that AI tools were perceived more strongly as supports for engagement, content relevance, and learning analytics than as clear drivers of academic performance, underscoring the need to separate AI-assisted activity from verified learning gain.

The concept of epistemic autonomy is central to the educational implications of AI integration. Epistemic autonomy refers to the capacity of individuals to evaluate evidence, form independent judgments, and maintain critical engagement with knowledge claims rather than deferring to authority or automated systems. Research on automation bias — the well-documented tendency of human operators to over-rely on automated system outputs even when those outputs are incorrect — suggests that this autonomy is not a stable default but a capacity that requires active cultivation and that is vulnerable to erosion under conditions of high AI availability and perceived AI reliability (Skitka et al., 1999). The behavioral implications are significant: as AI systems become more capable and more pervasively integrated into knowledge work, the conditions for epistemic autonomy may require deliberate institutional and pedagogical design rather than arising naturally from the cognitive environment.

The professional and organizational dimensions of AI-integrated cognition extend these concerns beyond formal education. In medicine, law, finance, and research, AI systems are increasingly functioning as epistemic partners in high-stakes judgment processes, and the behavioral literature on human-AI decision-making has documented a range of adaptation patterns. Cummings' work on human supervisory control of automated systems identifies a systematic tension between the efficiency gains that automation provides and the degradation of the human operator's situational awareness and independent judgment capacity over time — a dynamic she terms the irony of automation (Cummings, 2006). Applied to knowledge work more broadly, this dynamic raises the question of whether AI integration achieves its productivity gains at the cost of precisely the cognitive capacities that give human cognition its distinctive value. A parallel pattern is visible in accounting practice, where AI is more readily accepted for low-risk assistive functions than for core judgment tasks because traceability, verification burden, and accountable sign-off remain decisive constraints (Bendal et al., 2026).

The social dimensions of AI-integrated cognition add a further layer of complexity. The sociology of knowledge has long recognized that knowledge is not merely an individual cognitive achievement but a social institution, maintained through communities of practice, peer review, professional credentialing, and shared epistemic norms (Laurent, 2023; Rolin, 2017). The integration of AI into knowledge production potentially disrupts these social mechanisms in ways that go beyond individual cognitive adaptation. When AI systems generate research summaries, draft professional documents, and synthesize evidence, they may alter the social processes through which knowledge claims are validated, contested, and refined in ways that affect not only the quality of individual knowledge products but the collective epistemic health of the communities that produce and rely on them (Floridi, 2023).

Questions of epistemic trust — how human beings calibrate their confidence in knowledge sources — are particularly relevant here. Research on source monitoring and the social epistemology of testimony suggests that human epistemic trust is finely calibrated to social cues about the competence, honesty, and shared values of knowledge sources (Sperber et al., 2010). AI systems present a distinctive epistemic trust challenge because they do not provide the social cues that human trust calibration has evolved to rely on, yet they produce outputs with a surface fluency and apparent confidence that can activate trust responses inappropriately. The behavioral consequences of miscalibrated epistemic trust in AI outputs — including both under-trust that wastes the epistemic value AI can provide

and over-trust that allows AI errors to propagate unchecked — are among the most practically significant implications of AI integration and among the least theoretically developed in the current literature.

2.6 Synthesis of Literature

The five thematic domains surveyed in this review converge on a set of conclusions that are individually suggestive and collectively significant. Taken together, they establish that the emergence of machine learning constitutes not merely a technological development but a genuine epistemological and behavioral event — one that the existing literature has begun to recognize but has not yet fully theorized in an integrated way.

The first convergent pattern is the insufficiency of classical epistemological frameworks as descriptive or evaluative tools for ML. Across the philosophical literature, from internalist accounts of justified belief to phenomenological accounts of intentional understanding, the consistent finding is that ML systems produce outputs that function epistemically — that is, that serve as the basis for human belief, judgment, and action — through processes that satisfy none of the classical criteria for genuine knowing. The second convergent pattern is that human cognition is already behaviorally reorganizing around AI integration, following patterns documented in the cognitive offloading and extended mind literatures. The third convergent pattern is that the behavioral consequences of AI integration are double-edged: the same offloading that can free reasoning for higher-order tasks can also erode intrinsic capacities that deep engagement with difficulty develops. The fourth convergent pattern is that the social and institutional dimensions of AI integration are receiving less scholarly attention than the individual cognitive dimensions, despite their arguably greater practical significance.

2.7 Gaps in the Literature

Despite the breadth and depth of the literatures surveyed, significant gaps remain. The most consequential is the absence of integrative theoretical frameworks that bring the philosophical, cognitive-scientific, and behavioral literatures into direct conversation. A second significant gap concerns temporal scope: the existing literature is largely synchronic, documenting current adaptations without addressing long-term developmental consequences. A third gap concerns the social epistemology dimensions of AI integration — the mechanisms through which AI-generated knowledge claims enter and circulate within social epistemic communities. A fourth gap is the undertheorization of epistemic identity — the dimension of self-understanding related to individuals' sense of themselves as knowing agents. Finally, there is a methodological gap: the qualitative dimensions of human experience in AI-integrated cognitive environments have received relatively little attention compared to behavioral and performance outcomes.

2.8 Contribution of the Present Paper

The present paper makes several contributions to the literature that collectively address the gaps identified above. Its primary contribution is integrative: by bringing classical epistemology, extended mind theory, evolutionary epistemology, and the behavioral literature on AI integration into direct analytical dialogue, this paper constructs a framework that none of these literatures has developed individually. Its second contribution is conceptual clarification, sharpening the distinction between ML as a technical artifact and ML as an epistemic phenomenon. Its third contribution is a theoretically grounded framework for understanding the behavioral consequences of AI integration that goes beyond the current literature's largely reactive and domain-specific findings. Its fourth contribution is to place the social epistemology dimensions of AI integration within a theoretical framework that gives them their appropriate weight.

3. Methodology

3.1 Research Design

This paper adopts a philosophical-psychological conceptual synthesis as its primary research design. It does not report the findings of an empirical study, administer instruments to participants, collect primary qualitative data, or conduct fieldwork. Instead, it develops an original analytical argument through the systematic integration of theoretical frameworks and scholarly literature drawn from philosophy of mind, cognitive science, evolutionary epistemology, and behavioral research on human-AI interaction. This design is appropriate when the primary scholarly contribution consists of conceptual clarification, theoretical integration, and the identification of analytical frameworks that the existing literature has not developed — conditions that characterize the present inquiry.

Conceptual synthesis as a research design has an established tradition within the behavioral and social sciences (Paradice, 2010; Yazdani et al., 2021). It is distinct from both narrative review, which summarizes existing literature (Ghosh & Choudhury, 2025), and from systematic review, which follows a protocol-driven procedure for exhaustive literature retrieval and appraisal (Grant & Booth, 2009). Conceptual synthesis is instead characterized by the use of literature as intellectual material for constructing original theoretical positions, mapping conceptual relationships, resolving apparent contradictions across disciplinary traditions, and identifying productive directions for future empirical inquiry (Torraco, 2016).

3.2 Source Orientation and Inclusion Logic

The paper draws on four primary bodies of literature, each corresponding to one of the thematic domains established in the literature review. These are the literature on philosophy of mind and classical epistemology; artificial intelligence and philosophy of AI; cognitive science on extended mind theory and cognitive offloading; and behavioral and educational research on human-AI integration. Within each domain, source selection prioritized theoretical relevance rather than exhaustiveness. Primary theoretical sources — including foundational philosophical texts by Aristotle, Descartes, Kant, Husserl, and Heidegger — were included because the paper's central argument required an accurate characterization of the classical tradition against which ML is being assessed. Empirical sources from cognitive science and behavioral research were included when they provided evidence bearing on the theoretical claims the paper develops.

3.3 Thematic Grouping and Analytical Procedure

The five literature themes were developed through an iterative process of conceptual mapping, which used the central research problem as an organizing principle for identifying the relevant disciplinary domains and their internal debates. The themes followed a deliberate logical sequence: from the foundational philosophical tradition that establishes the baseline for comparison, through the epistemological analysis of ML itself, through the cognitive scientific frameworks that connect philosophical analysis to behavioral evidence, through the evolutionary dimension that provides temporal depth, to the behavioral and educational literature that grounds the argument in observable human consequences. Within each thematic section, the analytical procedure followed a consistent pattern: establishing foundational claims, identifying major agreements and disagreements, assessing contributions to the paper's central argument, and noting what remains unresolved.

3.4 Interpretive Framework

The paper's interpretive framework is explicitly pluralist, drawing on multiple philosophical and theoretical traditions rather than committing to a single school of thought as a master framework. Classical epistemology provides the baseline and the vocabulary for identifying what is epistemologically significant about ML's divergence from the tradition. Extended mind theory provides the framework for understanding ML integration as a behavioral and cognitive phenomenon. Evolutionary epistemology provides the temporal and adaptive perspective. Behavioral and educational research provides empirical texture and social relevance. These frameworks are not always fully compatible with one another — there are genuine tensions, for instance, between the phenomenological tradition's insistence on intentional subjectivity as constitutive of knowing and the extended mind thesis's functionalist criteria for cognitive extension (Meacham, 2017). The paper does not resolve these tensions artificially but uses them as productive sites of analysis.

3.5 Limitations

Several limitations of the present approach should be acknowledged. As a conceptual paper, it does not generate primary empirical evidence, and cannot therefore make claims about how specific individuals or populations actually experience, respond to, or are affected by AI integration. The paper also primarily orients its scope toward the Western philosophical tradition, and does not engage systematically and deeply with non-Western epistemological traditions. Additionally, the pace of development in both AI research and the behavioral literature on AI integration means that more recent findings may supersede some specific empirical claims. Finally, the evolutionary epistemology arguments, while grounded in an established scholarly tradition, involve a degree of extrapolation from current conditions to longer-term developmental trajectories that cannot be empirically verified.

4. Analytical Synthesis and Discussion

4.1 The Epistemological Inversion Thesis

The central analytical finding of this paper can be precisely stated: machine learning does not extend the classical epistemological tradition — it structurally inverts it. This inversion is not a matter of degree but of kind, and understanding its precise dimensions is necessary before any productive assessment of its implications for human cognition can be made.

The classical tradition, across its major variations, proceeds epistemologically from the general to the particular (Ninos, 2024). Aristotle's syllogistic moves from universal premises to particular conclusions (Raab, 2018). Descartes' method moves from foundational certainties, established through rational intuition, to derived knowledge of the world (Dellsén, 2017). Kant's transcendental epistemology moves from the a priori categories of the understanding to the organized experience those categories make possible (Kant, 1998). In each case, the epistemic process is directional, intentional, and anchored in rational principles that precede and organize the encounter with particular instances (Stroud, 2019). The knower, in the classical picture, brings something essential to the encounter with the world — structure, categories, principles, rational norms — that is not itself derived from the particular data being processed (Veldman & Swagerman, 2018).

Machine learning inverts this structure at every point. The deep learning process begins with particular instances — vast quantities of them — and produces generalizations, representations, and functional abstractions through an optimization procedure that adjusts parameters to minimize prediction error across the training distribution (Bengio et al., 2018; Karaca, 2021). There are no prior principles, no rational norms guiding the process, no intentional subject organizing the encounter with data. What emerges from training is not a set of explicit rules or logical relationships but a distributed pattern of numerical weights whose functional properties constitute the system's representational capacity. The movement is from particulars to emergent pattern rather than from principles to conclusions — and crucially, this movement occurs without any of the intentional, phenomenological, or rationally transparent features that the classical tradition has treated as constitutive of knowing.

This inversion has three analytically distinct dimensions that deserve separate treatment. The first is the inversion of directionality: classical knowing moves top-down from principles to instances, while ML moves bottom-up from instances to emergent representations. The second is the inversion of transparency: classical knowing, in its paradigm cases, is in principle fully articulable — the inferential chain can be traced, the justification can be stated, the reasoning can be examined. ML outputs, particularly in deep learning, are in principle opaque — the functional representations that produce them are distributed across billions of parameters in ways that resist straightforward interpretive decomposition (Qamar & Bawany, 2023; Teng et al., 2022). The third is the inversion of the subject-object relationship: classical knowing presupposes a rational subject whose cognitive activity is the source of epistemic order, while ML produces epistemic outputs without any subject in the relevant sense present in the process.

What follows from this inversion is not (the paper argues) that ML fails to be epistemologically significant. It follows, rather, that its epistemological significance is of a different kind from what the tradition's vocabulary can capture. The outputs of ML systems function epistemically in the sense that they enter human cognitive processes and serve as the basis for belief, judgment, and action. What the inversion thesis establishes is that this functional role is achieved through means that the classical tradition did not anticipate and cannot straightforwardly evaluate, and that new conceptual frameworks are therefore required to understand what kind of epistemic resource ML represents and what its integration into human cognitive life means.

4.2 Machine Learning as the Most Radical Instance of Cognitive Extension

The extended mind thesis, applied to machine learning, yields an interpretation that is both theoretically productive and behaviorally consequential. If the criterion for genuine cognitive extension is functional integration — the reliable, accessible, and automatically endorsed role of an external resource in a cognitive process — then ML systems present a stronger case for cognitive extension than any previous cognitive tool, and simultaneously a case that is more philosophically complex than the thesis's original formulations anticipated.

Previous instances of cognitive extension — writing, mathematical notation, mechanical calculation, digital computation — shared an important property: their operations were, at least in principle, fully transparent to the human cognitive system that incorporated them (Wheeler, 2018). A written text preserves information in a form that the human reader can access, evaluate, and critically engage with. A calculator performs operations whose logical structure is fully intelligible even to a user who does not understand the electronic implementation. The human

cognitive system extending itself through these tools retains, in principle, full epistemic access to what the tool is doing — the extension is of cognitive capacity, not of cognitive opacity.

ML integration introduces opacity into the extended cognitive system in a way that has no clear precedent (Koskinen, 2023; Naeem & Hauser, 2024). When a physician uses an AI diagnostic system, a researcher uses an AI literature synthesis tool, or a student uses an AI reasoning assistant, the extended cognitive system incorporates a component whose operations are not transparent to the human component of the system. The human cannot, in general, trace the inferential path by which the AI component reached its output. The extended system as a whole therefore has representational capacities that exceed the epistemic access of its human component — a situation that is qualitatively different from all previous forms of cognitive extension.

This analysis generates a key distinction that the current literature has not articulated with sufficient clarity: the difference between capacity extension and opacity extension. Prior cognitive tools extended human cognitive capacity while preserving epistemic transparency (Piredda & Francesco, 2020; Wheeler, 2018). ML tools extend human cognitive capacity while introducing epistemic opacity (Barelli et al., 2024; Hayes et al., 2022). This distinction matters significantly for how AI integration should be evaluated, designed, and governed — not only philosophically but behaviorally and institutionally. An extended cognitive system that incorporates opacity is one in which the human component may lose the ability to evaluate the reliability of the system's outputs, to identify the conditions under which it fails, or to maintain the independent cognitive competence that would be necessary to function without it.

The behavioral literature on cognitive offloading and automation bias documents precisely the consequences of this opacity extension in practice (Bauer et al., 2023; Berberian et al., 2017). The well-documented tendency toward automation bias is not merely a failure of critical thinking by individual users (Bauer et al., 2023). It is a predictable behavioral consequence of incorporating an opaque, high-performing component into a cognitive system: when the opaque component's outputs are reliably correct in the vast majority of cases, the rational response is to trust them, and the behavioral cost of that trust only becomes apparent in the subset of cases where the component fails (Berberian et al., 2017; Bauer et al., 2023). The epistemological and behavioral problem is therefore structural rather than individual — it is a property of the relationship between human cognition and opaque AI components, not a correctable trait of particular users (Berberian et al., 2017; Bauer et al., 2023).

4.3 Evolutionary Epistemology and the Adaptive Significance of the ML Transition

The evolutionary epistemology framework adds a temporal and adaptive dimension to the analysis that fundamentally alters its significance. Viewed through this lens, the current moment of AI integration is not best understood as a discrete technological event but as a transition point in the long adaptive history of human cognitive ecology.

Plotkin's hierarchical account of adaptive systems, which distinguishes between genetic evolution, individual learning, and cultural transmission as nested adaptive processes operating on different timescales, provides a useful analytical structure for this interpretation (Jan, 2022). Human beings are biologically adaptive organisms whose genetic endowment includes cognitive capacities — perception, memory, inference, language — that evolved under selection pressures operating over hundreds of thousands of years (Pinker, 2010; Taylor et al., 2021). Layered on this biological foundation is a system of individual learning, and layered on it is a system of cultural transmission that allows adaptations to accumulate and propagate across generations (Jan, 2022). Machine learning is best understood as a new class of artifact within the cultural adaptive system — one whose properties are sufficiently novel to warrant treating it as a qualitatively distinct phase, rather than merely an incremental addition.

The adaptive logic of cognitive offloading suggests that human beings will, under conditions of high AI availability and reliability, progressively redistribute cognitive effort away from the functions that AI performs competently and toward those where human characteristics — intentionality, phenomenological engagement, contextual judgment, ethical reasoning — provide comparative advantage (Chirayath et al., 2025; Langlois, 2003). This redistribution is not a sign of cognitive degradation but of rational adaptive response to a changed environment. Writing did not destroy human memory — it transformed the role of memory within a restructured cognitive ecology (Heersmink, 2020; Morais & Kolinsky, 2020). Calculation tools did not destroy mathematical reasoning — they freed human attention for problems that exceed computational capacity (Misfeldt et al., 2015; Robinson & Burns, 2009).

The adaptive risk, however, is real and should not be minimized. Adaptive redistribution of cognitive effort requires that the functions being offloaded remain available for reactivation when the tool is unavailable or unreliable,

and that the human component of the extended cognitive system retains sufficient competence to evaluate the tool's outputs critically (Chen & de Beeck, 2021; O'Halloran et al., 2015). The evolutionary literature on niche construction suggests that the cognitive niche human beings are constructing through AI integration may, over sufficient time, alter the selection landscape for cognitive capacities in ways that cannot be reversed by individual choice or institutional policy (Atã & Queiroz, 2016; Pinker, 2010). This is a claim that cannot be empirically verified with current evidence, but it is a theoretically grounded concern that the scholarly literature has not yet taken appropriately seriously.

4.4 Behavioral and Social Consequences: Toward an Integrated Framework

The analytical threads developed in the preceding sections converge on a set of behavioral and social implications that, taken together, constitute the practical significance of the epistemological transition this paper describes.

At the individual level, the integration of ML into cognitive practice is producing a new form of epistemic identity challenge (Margondai et al., 2025; Richter & Schaller, 2025). Human beings have historically understood themselves, in part, through their cognitive capacities — their ability to reason, to know, to judge, and to create. As ML systems perform these functions with increasing competence (Jussupow et al., 2022; Waefler & Schmid, 2021), individuals in knowledge-intensive fields are confronted with questions about the nature and value of their own cognitive contributions that the existing psychological literature on identity and self-concept has not adequately addressed (Margondai et al., 2025). The educational literature's documentation of skill atrophy risk under conditions of high AI assistance points toward a behavioral correlate of this identity challenge: the gradual erosion of the intrinsic cognitive capacities that have constituted the basis of professional identity and self-efficacy in knowledge workers (Ganuthula, 2024; Kabashkin, 2025).

At the institutional level, the social mechanisms through which knowledge is validated, contested, and refined are under pressure from AI integration in ways that the sociology of knowledge has begun to document but not yet fully theorize (Collins, 2024; Deranty & Corbin, 2022; Gözütok, 2025). Peer review, professional credentialing, and communities of epistemic practice depend on the assumption that those participating in them bring genuine cognitive competence and independent judgment to the process (Gözütok, 2025; Hosseini & Horbach, 2023). As AI systems become capable of producing outputs that are indistinguishable from those of human experts in many surface respects, the social mechanisms for distinguishing genuine competence from AI-assisted performance face challenges that they were not designed to address (Gözütok, 2025; Mann et al., 2025).

At the societal level, the distribution of epistemic trust — the social allocation of credibility across individuals, institutions, and knowledge sources — is being reorganized in ways whose long-term consequences are not yet clear (Kheokao et al., 2025; Younas & Zeng, 2024). The combination of highly capable AI systems, widespread access, and the documented tendency toward automation bias creates conditions in which epistemic authority may become concentrated around AI outputs in ways that reduce the diversity of knowledge-producing perspectives that have historically been a source of epistemic resilience in human societies (Giovanni, 2025; Kheokao et al., 2025; Younas & Zeng, 2024).

Together, these analytical threads support the paper's central claim: that the emergence of machine learning constitutes an epistemological and behavioral event of sufficient significance to require integrated theoretical attention across the disciplines that study human knowing, and that the scholarly literature's current fragmentation across these disciplines is leaving the most important questions inadequately addressed.

5. Discussion of Findings and Implications

5.1 Theoretical Significance of the Epistemological Inversion

The epistemological inversion thesis developed in this paper carries theoretical implications that extend beyond the specific question of how machine learning relates to classical epistemology. At its deepest level, the inversion thesis suggests that human beings have produced, for the first time in their intellectual history, an artifact that performs epistemic functions through means that are structurally discontinuous with the cognitive processes through which those same functions are performed biologically. This discontinuity forces a separation between two dimensions of knowing that the classical tradition has treated as inseparable: the epistemic process and the epistemic product. Machine learning makes this coupling untenable as a universal criterion, because it produces epistemically functional

outputs through processes that satisfy none of the conditions the tradition has required — and in doing so, it compels a revision of the tradition's foundational assumptions that is philosophically productive rather than merely critical.

5.2 Implications for Human Identity and Epistemic Agency

Perhaps the most consequential implication of the paper's findings for contemporary human life concerns the relationship between cognitive capacity and identity. The behavioral literature reviewed in this paper documents a consistent pattern: as AI systems perform functions previously understood as distinctly human, individuals in knowledge-intensive fields experience challenges to their sense of professional identity, intellectual agency, and cognitive self-efficacy (Jussupow et al., 2022; Zhu et al., 2025). This pattern reflects a deep connection between human self-understanding and the experience of knowing — a connection the philosophical tradition from Descartes onward has treated as constitutive of human dignity (Erk, 2010; Harman, 2020). When ML systems perform these functions with competence that rivals or exceeds human performance, the implicit connection between knowing and human distinctiveness exerts pressure resulting in psychological consequences that the behavioral literature is only beginning to document (Riley et al., 2025).

If epistemic identity — the sense of oneself as a capable and independent knowing agent — is a psychologically significant dimension of human self-concept (Demerath, 2006; Osbeck & Nersessian, 2017), then the conditions under which it is maintained or eroded under AI integration are matters of genuine behavioral concern. The finding that AI assistance can produce skill atrophy in domains where productive cognitive struggle is itself a mechanism of development (Kabashkin, 2025; Macnamara et al., 2024) suggests that the preservation of epistemic agency may require deliberate pedagogical and institutional design, prioritizing the development and maintenance of independent cognitive competence not merely for instrumental reasons but because that competence is constitutive of a form of human flourishing that cannot be replaced by reliable AI outputs.

5.3 Implications for Epistemic Trust and Social Knowledge

The social epistemological implications of the paper's findings are significant and underappreciated in the existing literature. Human epistemic communities — scientific communities, professional bodies, educational institutions, democratic publics — function through complex social mechanisms of trust, contestation, and validation that have evolved over centuries to manage the collective production and circulation of knowledge (Lucio-Arias & Leydesdorff, 2009; Roth & Cointet, 2009). The behavioral dynamics documented in the paper — automation bias, cognitive offloading, opacity extension — introduce systemic vulnerabilities into these social mechanisms that operate at a level for which they were not designed. When individual participants in epistemic communities increasingly rely on opaque AI systems for the production of knowledge claims, the social process of peer review, critical contestation, and expert evaluation faces a new challenge: distinguishing genuine independent judgment from AI-assisted outputs that mimic their surface features.

The epistemic trust literature's findings regarding the mismatch between human trust-calibration mechanisms and AI output characteristics further deepen this concern (Hannig et al., 2025; Sargeant et al., 2025). The combination of trust-calibration mismatch with the high surface fluency of modern ML outputs creates conditions in which epistemic trust may be systematically miscalibrated at scale — a form of collective epistemic vulnerability whose consequences for democratic deliberation, scientific integrity, and professional accountability have not yet been adequately theorized.

5.4 Boundaries of Conclusion

Intellectual honesty requires that the boundaries of what this paper can and cannot conclude be stated explicitly. The epistemological inversion thesis is analytically defensible on the basis of the philosophical and cognitive scientific literature reviewed. The structural divergence between classical human knowing and ML processing is a matter of established fact rather than interpretation. Claims about long-term adaptive consequences are theoretically grounded but empirically unverifiable with current evidence, offered as analytically motivated possibilities rather than established conclusions. The behavioral claims are grounded in existing empirical literature, though this literature is limited by its recency and the pace of change in AI capabilities. What the paper can conclude with confidence is that the emergence of machine learning raises epistemological and behavioral questions of genuine significance that the existing scholarly literature has not yet addressed with the integration and analytical depth they deserve.

6. Conclusions and Recommendations

6.1 Summary of Principal Findings

This paper set out to address a question at the intersection of philosophy, cognitive science, and behavioral research: what is the epistemological and behavioral significance of machine learning for human cognition, and how should the relationship between ML and the classical tradition of human knowing be theorized? The analysis developed across the preceding sections supports several conclusions that, taken together, constitute a coherent and analytically grounded response.

The first and most foundational conclusion is that machine learning constitutes a genuine epistemological rupture with the classical tradition rather than an extension or amplification of it. The structural inversion the paper identifies — from principle-to-conclusion to data-to-emergent-pattern, from intentional to nonintentional process, from transparent to opaque inference — is not a matter of degree but of kind. The second conclusion is that the extended mind framework reveals a qualitatively new form of cognitive extension that introduces opacity into the human cognitive system in a way that prior cognitive tools did not. The third conclusion is that evolutionary epistemology provides the most adequate available framework for understanding the long-term adaptive significance of ML integration. The fourth conclusion is that the social and institutional dimensions of AI integration are receiving inadequate scholarly attention relative to their practical significance.

6.2 Recommendations for Research

The gaps identified in this paper point to several specific directions for future scholarly inquiry. The most immediate priority is the development of integrative empirical studies that bring the philosophical dimensions of the epistemological inversion thesis into dialogue with behavioral evidence about how specific populations are actually experiencing and adapting to deep AI integration. The qualitative dimensions of this experience are particularly underexplored — how human beings narrate, interpret, and make meaning of their cognitive collaboration with AI systems, what the phenomenology of knowing with a machine feels like, and how this experience shapes epistemic identity and agency are questions that require sustained qualitative investigation.

A second priority is the longitudinal investigation of cognitive offloading effects under conditions of sustained AI integration. Understanding whether skill atrophy effects are reversible, cumulative, or domain-specific requires longitudinal designs that track cognitive development across extended periods of AI use. A third priority is the development of social epistemological research on the collective consequences of AI integration for epistemic communities — how AI-generated knowledge claims circulate, how they affect epistemic trust and validation dynamics, and what institutional designs are most effective at preserving epistemic independence under conditions of high AI availability.

6.3 Recommendations for Practice

The paper's findings carry several implications for educational and institutional practice. For educational design, the findings support a shift from efficiency-oriented AI integration toward development-oriented AI integration calibrated to preserve the intrinsic cognitive capacities that deep engagement with difficulty develops. This means designing learning environments in which AI assistance is strategically scaffolded in ways that ensure productive cognitive struggle remains a feature of the learning experience. Comparable systems-level pressures in health professions education suggest that digital-era learning challenges should be interpreted alongside generational expectations, faculty workload, institutional viability, and governance capacity rather than reduced to individual learner adaptation (Bermido et al., 2025).

For professional and organizational practice, the findings support the development of epistemic governance frameworks that maintain human cognitive competence and independent judgment as explicit institutional values. Organizations integrating AI into knowledge-intensive workflows should invest in deliberate mechanisms for preserving human cognitive capacities — including rotation of responsibilities, periodic unassisted performance requirements, and explicit cultivation of the critical evaluation skills that automation bias tends to erode (Branda & Ciccozzi, 2026; Sarkar et al., 2024). For policy and institutional design more broadly, the findings support greater investment in the social epistemological infrastructure that collective knowledge-making depends on, with specific attention to the new challenges that AI integration poses. From an organizational analytics perspective, Atento, Quinto, Espelita, and Castaneda (2025) similarly frame effective analytics as a socio-technical pathway that links data integration, analytics capability, decision quality, and aligned outcomes, a useful parallel for treating AI governance as more than technical deployment.

6.4 Closing Reflection

The question this paper has pursued — what machine learning means for human knowing — is not merely a philosophical or technical question. It is a question about the kind of cognitive beings human beings will be in the decades and centuries ahead, and about the conditions under which the distinctively human dimensions of knowing — intentionality, phenomenological engagement, rational accountability, epistemic agency — will be preserved, developed, or allowed to atrophy. The framework this paper develops does not answer that question, but it provides the analytical foundation from which an answer can begin to be built. That building is, the paper argues, among the most important intellectual tasks of the present moment — and it requires the full collaborative engagement of philosophy, cognitive science, behavioral research, and the social sciences that study the human conditions of knowledge.

References

Adams, F., & Aizawa, K. (2001). The bounds of cognition. Philosophical Psychology, 14(1), 43. https://doi.org/10.1080/09515080120033571

Agarwal, A. (2017). Knowing "knowledge" and "to know": An overview of concepts. International Journal of Research-GRANTHAALAYAH, 5(11), 86. https://doi.org/10.29121/granthaalayah.v5.i11.2017.2331

Aithal, P. S. (2023). Super-intelligent machines: Analysis of developmental challenges and predicted negative consequences. International Journal of Applied Engineering and Management Letters, 109. https://doi.org/10.47992/ijaeml.2581.7000.0191

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A. Q., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 8(1). https://doi.org/10.1186/s40537-021-00444-8

Andrews, M. (2025). The immortal science of ML: Machine learning and the theory-free ideal. Erkenntnis. https://doi.org/10.1007/s10670-025-01010-x

Aristotle. (1994). Posterior analytics (J. Barnes, Trans.). Clarendon Press. (Original work published c. 350 BCE)

Atã, P., & Queiroz, J. (2016). Habit in semiosis: Two different perspectives based on hierarchical multi-level system modeling and niche construction theory. In Studies in applied philosophy, epistemology and rational ethics (p. 109). Springer Nature. https://doi.org/10.1007/978-3-319-45920-2_7

Atento, R. G. (2025). Exploring e-learning for sustainable development: Integrating SDGs in management education at Philippine higher education institutions. International Journal of Health & Business Analytics, 1(1). https://doi.org/10.65166/2qcjx561

Atento, R. G., Quinto, L., Espelita, C. A. M., & Castaneda, C. (2025). Integrating business and health analytics: A conceptual framework for dual outcomes in healthcare. International Journal of Health & Business Analytics, 1(1). https://doi.org/10.65166/04pdc866

Atento, R. G. O., Quinto, L. F., Espelita, C. A. M., & San Juan, F. M. (2025). Narrative health analytics: Integrating empathy, data, and ethics in patient-centered healthcare. International Journal of Health and Business Analytics, 1(2), 1-33. https://doi.org/10.65166/yxgx8e59

Avello, D., & Zurita, S. (2025). Exploring the nexus of academic integrity and artificial intelligence in higher education: A bibliometric analysis. International Journal for Educational Integrity, 21(1). https://doi.org/10.1007/s40979-025-00199-2

Barelli, E., Lodi, M., Branchetti, L., & Levrini, O. (2024). Epistemic insights as design principles for a teaching-learning module on artificial intelligence. Science & Education. https://doi.org/10.1007/s11191-024-00504-4

Bauer, K., Zahn, M. von, & Hinz, O. (2023). Please take over: XAI, delegation of authority, and domain knowledge. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4512594

Bendal, A., Sabasa, S. A., Espelita, C. A. M. H., & Atento, R. G. O. (2026). Artificial intelligence as disruptive technology in accounting: A qualitative study of practitioner perceptions on automation, judgment, and

decision support. Journal of Enterprise Strategy & Management Innovation, 1(1). https://doi.org/10.65166/0sdayg70

Bengio, Y., Lodi, A., & Prouvost, A. (2018). Machine learning for combinatorial optimization: A methodological tour d'horizon. arXiv. https://doi.org/10.48550/arxiv.1811.06128

Berberian, B., Somon, B., Sahaï, A., & Gouraud, J. (2017). The out-of-the-loop brain: A neuroergonomic approach of the human automation interaction. Annual Reviews in Control, 44, 303. https://doi.org/10.1016/j.arcontrol.2017.09.010

Bermido, C. M., Quinto, L. F., & Atento, R. G. O. (2025). A qualitative thematic review of contemporary challenges affecting health professions education: Implications for higher education leadership. International Journal of Health and Business Analytics, 1(2). https://doi.org/10.65166/yfm5w791

Black, R. W., & Tomlinson, B. (2025). University students describe how they adopt AI for writing and research in a general education course. Scientific Reports, 15(1), 8799. https://doi.org/10.1038/s41598-025-92937-2

Block, J., & Kuckertz, A. (2024). What is the future of human-generated systematic literature reviews in an age of artificial intelligence? Management Review Quarterly. https://doi.org/10.1007/s11301-024-00471-8

Bond, E. (2021). Archaeology of human consciousness: An integrated narrative of cognitive evolution. Advances in Anthropology, 11(3), 201. https://doi.org/10.4236/aa.2021.113013

Botha, Griffiths, D., & Prozesky, M. (2021). Epistemological decolonization through a relational knowledge-making model. Africa Today, 67(4), 50. https://doi.org/10.2979/africatoday.67.4.04

Branda, F., & Ciccozzi, M. (2026). The comfort of automation: Why cognitive sovereignty matters in AI-driven life sciences. Artificial Intelligence in the Life Sciences, 9, 100158. https://doi.org/10.1016/j.ailsci.2026.100158

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society, 3(1). https://doi.org/10.1177/2053951715622512

Cassinadri, G., & Fasoli, M. (2023). Rejecting the extended cognition moral narrative: A critique of two normative arguments for extended cognition. Synthese, 202(5). https://doi.org/10.1007/s11229-023-04397-8

Chen, C., & de Beeck, H. O. (2021). Perceptual learning with complex objects: A comparison between full-practice training and memory reactivation. eNeuro, 8(2). https://doi.org/10.1523/eneuro.0008-19.2021

Chirayath, G., Premamalini, K., & Joseph, J. (2025). Cognitive offloading or cognitive overload? How AI alters the mental architecture of coping. Frontiers in Psychology, 16, 1699320. https://doi.org/10.3389/fpsyg.2025.1699320

Cibu, B., Crăciun, L., Molănescu, A. G., & Cotfas, L. (2025). Exploring the educational applications of large language models: A systematic review and topic analysis. Electronics, 14(23), 4683. https://doi.org/10.3390/electronics14234683

Clark, A. (2001). Natural-born cyborgs? In Lecture notes in computer science (p. 17). Springer. https://doi.org/10.1007/3-540-44617-6_2

Clark, A. (2003). Natural-born cyborgs: Minds, technologies, and the future of human intelligence. Oxford University Press.

Clark, A. (2025). Extending minds with generative AI. Nature Communications, 16(1), 4627. https://doi.org/10.1038/s41467-025-59906-9

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

Clark, A. J. (1986). Evolutionary epistemology and the scientific method. Philosophica, 37. https://doi.org/10.21825/philosophica.82528

Collins, H. (2024). Why artificial intelligence needs sociology of knowledge: Parts I and II. AI & Society, 40(3), 1249. https://doi.org/10.1007/s00146-024-01954-8

Cottingham, J. (Trans.). (2016). Meditations on first philosophy (R. Descartes). Routledge. https://doi.org/10.4324/9781315508818-8

Cummings, M. L. (2006). Automation and accountability in decision support system interface design. The Journal of Technology Studies, 32(1), 23. https://doi.org/10.21061/jots.v32i1.a.4

Dellsén, F. (2017). Certainty and explanation in Descartes's philosophy of science. HOPOS: The Journal of the International Society for the History of Philosophy of Science, 7(2), 302. https://doi.org/10.1086/692013

Demerath, L. (2006). Epistemological identity theory: Reconceptualizing commitment as self-knowledge. Sociological Spectrum, 26(5), 491. https://doi.org/10.1080/02732170600786208

Dennett, D. C. (1995). Darwin's dangerous idea: Evolution and the meanings of life. Simon & Schuster.

Deranty, J., & Corbin, T. (2022). Artificial intelligence and work: A critical review of recent research from the social sciences. AI & Society, 39(2), 675. https://doi.org/10.1007/s00146-022-01496-x

Dergaa, I., Saad, H. B., Glenn, J. M., Amamou, B., Aissa, M. B., Guelmami, N., Fekih-Romdhane, F., & Chamari, K. (2024). From tools to threats: A reflection on the impact of artificial-intelligence chatbots on cognitive health. Frontiers in Psychology, 15, 1259845. https://doi.org/10.3389/fpsyg.2024.1259845

Diržytė, A. (2025). Large language models and the enhancement of human cognition: Some theoretical insights. Filosofija Sociologija, 36(1). https://doi.org/10.6001/fil-soc.2025.36.1.2

Dreyfus, H. L. (1978). What computers can't do: The limits of artificial intelligence. Harper & Row.

Embracing the ubiquity of machines. (2024). Nature Human Behaviour, 8(10), 1823. https://doi.org/10.1038/s41562-024-02049-6

Erk, C. (2010). Health, rights and dignity: Philosophical reflections on an alleged human right. https://doi.org/10.26530/oapen_626362

Floridi, L. (2023). The ethics of artificial intelligence. Oxford University Press. https://doi.org/10.1093/oso/9780198883098.001.0001

Ganuthula, V. R. R. (2024). The paradox of augmentation: A theoretical model of AI-induced skill atrophy. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4974044

Garzón, J., Patiño, E., & Marulanda, C. (2025). Systematic review of artificial intelligence in education: Trends, benefits, and challenges. Multimodal Technologies and Interaction, 9(8), 84. https://doi.org/10.3390/mti9080084

George, A. S., Baskar, T., & Srikaanth, P. B. (2024). The erosion of cognitive skills in the technological age: How reliance on technology impacts critical thinking, problem-solving, and creativity. Zenodo. https://doi.org/10.5281/zenodo.11671150

Ghosh, A., & Choudhury, S. (2025). Understanding different types of review articles: A primer for early career researchers. Indian Journal of Psychiatry, 67(5), 535. https://doi.org/10.4103/indianjpsychiatry.indianjpsychiatry_373_25

Giovanni, A. (2025). Cyber humanism in education: Reclaiming agency through AI and learning sciences. arXiv. https://doi.org/10.48550/arxiv.2512.16701

Gözütok, T. T. (2025). Yapay zekânın akademik yayın hakemliğindeki rolü üzerine epistemik ve etik bir sorgulama. DergiPark. https://doi.org/10.20981/kaygi.1704390

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. Health Information & Libraries Journal, 26(2), 91. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Grinschgl, S., Meyerhoff, H. S., Schwan, S., & Papenmeier, F. (2020). From metacognitive beliefs to strategy selection: Does fake performance feedback influence cognitive offloading? Psychological Research, 85(7), 2654. https://doi.org/10.1007/s00426-020-01435-9

Grinschgl, S., & Neubauer, A. C. (2022). Supporting cognition with modern technology: Distributed cognition today and in an AI-enhanced future. Frontiers in Artificial Intelligence, 5. https://doi.org/10.3389/frai.2022.908261

Gurevich, Y., & Blass, A. (2024). On logic and generative AI. arXiv. https://doi.org/10.48550/arxiv.2409.14465

Hannig, L., Bush, A., Aksoy, M., Trappen, T., Becker, S., & Ontrup, G. (2025). Campus AI vs. commercial AI: How customizations shape trust and usage of LLM as-a-service chatbots. arXiv. https://doi.org/10.48550/arxiv.2509.15826

Harman, G. (2020). The only exit from modern philosophy. Open Philosophy, 3(1), 132. https://doi.org/10.1515/opphil-2020-0009

Hatfield, G., & Goldman, A. I. (1989). Epistemology and cognition. The Philosophical Review, 98(3), 386. https://doi.org/10.2307/2185025

Hayes, P., van de Poel, I., & Steen, M. (2022). Moral transparency of and concerning algorithmic tools. AI and Ethics, 3(2), 585. https://doi.org/10.1007/s43681-022-00190-4

Heersmink, R. (2020). Narrative niche construction: Memory ecologies and distributed narrative identities. Biology & Philosophy, 35(5). https://doi.org/10.1007/s10539-020-09770-2

Heersmink, R., & Knight, S. (2018). Distributed learning: Educating and assessing extended cognitive systems. Philosophical Psychology, 31(6), 969. https://doi.org/10.1080/09515089.2018.1469122

Heidegger, M. (1962). Being and time (J. Macquarrie & E. Robinson, Trans.). Harper & Row. (Original work published 1927)

Hosseini, M., & Horbach, S. P. J. M. (2023). Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. Research Square. https://doi.org/10.21203/rs.3.rs-2587766/v1

Husserl, E. (1970). Logical investigations (J. N. Findlay, Trans.). Routledge. (Original work published 1900–1901)

Jan, S. (2022). Music in evolution and evolution in music. Open Book Publishers. https://doi.org/10.11647/obp.0301

Jensen, L. X., Buhl, A., Sharma, A. V. N. L., & Bearman, M. (2024). Generative AI and higher education: A review of claims from the first months of ChatGPT. Higher Education. https://doi.org/10.1007/s10734-024-01265-3

Jussupow, E., Spohrer, K., & Heinzl, A. (2022). Identity threats as a reason for resistance to artificial intelligence: Survey study with medical students and professionals. JMIR Formative Research, 6(3). https://doi.org/10.2196/28750

Kabashkin, I. (2025). Cognitive atrophy paradox of AI–human interaction: From cognitive growth and atrophy to balance. Information, 16(11), 1009. https://doi.org/10.3390/info16111009

Kant, I. (1998). Critique of pure reason (P. Guyer & A. W. Wood, Trans.). Cambridge University Press. (Original work published 1781) https://doi.org/10.1017/cbo9780511804649

Karaca, K. (2021). Values and inductive risk in machine learning modelling: The case of binary classification models. European Journal for Philosophy of Science, 11(4). https://doi.org/10.1007/s13194-021-00405-1

Kheokao, D., Kheokao, J., Nopsuwam, R., & Boonwattanopas, D. (2025). AI, sovereignty, and the reshaping of knowledge production and public opinion. Asian Journal of Political Science. https://doi.org/10.15206/ajpor.2025.13.4.513

Koskinen, I. (2023). We have no satisfactory social epistemology of AI-based science. Social Epistemology, 38(4), 458. https://doi.org/10.1080/02691728.2023.2286253

Krishnan, G., Singh, S., Pathania, M., Gosavi, S., Abhishek, S., Parchani, A., & Dhar, M. (2023). Artificial intelligence in clinical medicine: Catalyzing a sustainable global healthcare paradigm. Frontiers in Artificial Intelligence, 6. https://doi.org/10.3389/frai.2023.1227091

Langlois, R. N. (2003). Cognitive comparative advantage and the organization of work: Lessons from Herbert Simon's vision of the future. Journal of Economic Psychology, 24(2), 167. https://doi.org/10.1016/s0167-4870(02)00201-5

Larsen, B. C. (2022). Governing artificial intelligence: Lessons from the United States and China. Danish Institute for International Studies.

Laurent, B. (2023). Institutions of expert judgment: The production and use of objectivity in public expertise. In Oxford handbook of expertise (p. 214). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190848927.013.10

LeCun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep learning. Nature, 521(7553), 436. https://doi.org/10.1038/nature14539

Liu, D., Fan, G. F., & Pan, L. (2026). Tool, tutor, or crutch? A grounded theory of cognitive scaffolding and offloading in AI-assisted programming education. International Journal of STEM Education, 13(1). https://doi.org/10.1186/s40594-025-00592-w

Lucio-Arias, D., & Leydesdorff, L. (2009). The dynamics of exchanges and references among scientific texts, and the autopoiesis of discursive knowledge. Journal of Informetrics, 3(3), 261. https://doi.org/10.1016/j.joi.2009.03.003

Lykhatskyi, A. (2025). Hybrid epistemology: Emergent knowledge forms in the age of human-AI cognitive integration. The Bulletin of Yaroslav Mudryi National Law University, 3(66). https://doi.org/10.21564/2663-5704.66.337968

Maclure, J. (2021). AI, explainability and public reason: The argument from the limitations of the human mind. Minds and Machines, 31(3), 421. https://doi.org/10.1007/s11023-021-09570-x

Macnamara, B. N., Berber, I., Çavuşoğlu, M. C., Krupinski, E. A., Nallapareddy, N., Nelson, N. E., Smith, P. J., Wilson-Delfosse, A. L., & Ray, S. (2024). Does using artificial intelligence assistance accelerate skill decay and hinder skill development without performers' awareness? Cognitive Research: Principles and Implications, 9(1). https://doi.org/10.1186/s41235-024-00572-8

Mai, X., Zeng, T., Lin, J., Wang, H., Chang, Y., Kang, Y., Wang, Y., & Zhang, W. (2024). From efficient multimodal models to world models: A survey. arXiv. https://doi.org/10.48550/arxiv.2407.00118

Mann, S. P., Aboy, M., Seah, J. J., Lin, Z., Luo, X., Rodger, D., Zohny, H., Minssen, T., Savulescu, J., & Earp, B. D. (2025). AI and the future of academic peer review. arXiv. https://doi.org/10.48550/arxiv.2509.14189

Marcus, G., & Davis, E. (2019). Rebooting AI: Building artificial intelligence we can trust. Pantheon Books.

Margondai, A., Willox, S., & Mouloua, M. (2025). Autonomy in transition: AI, self-identity, and the evolution of human agency. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 69(1), 1403. https://doi.org/10.1177/10711813251364790

Meacham, D. (2017). Introduction: Critiquing technologies of the mind: Enhancement, alteration, and anthropotechnology. Phenomenology and the Cognitive Sciences, 16(1), 1. https://doi.org/10.1007/s11097-017-9505-3

Misfeldt, M., Barbin, É., Jankvist, U. T., & Kjeldsen, T. H. (2015). Panel debate: Technics and technology in mathematics and mathematics education. Research Portal Denmark.

Mollick, E., & Mollick, L. (2023). Assigning AI: Seven approaches for students, with prompts. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4475995

Morais, J., & Kolinsky, R. (2020). Seeing thought: A cultural cognitive tool. Journal of Cultural Cognitive Science, 5(2), 181. https://doi.org/10.1007/s41809-020-00059-0

Naeem, H., & Hauser, J. (2024). Should we discourage AI extension? Epistemic responsibility and AI. Philosophy & Technology, 37(3). https://doi.org/10.1007/s13347-024-00774-4

Natali, C., Marconi, L., Duran, L. D. D., & Cabitza, F. (2025). AI-induced deskilling in medicine: A mixed-method review and research agenda for healthcare and beyond. Artificial Intelligence Review, 58(11). https://doi.org/10.1007/s10462-025-11352-1

Ninos, G. (2024). Hegel's theory of finite cognition and Marx's critique of political economy. Hegel Bulletin, 1. https://doi.org/10.1017/hgl.2024.22

Oermann, M. H., Owens, J. K., Carter-Templeton, H., Peterson, G. M., & Bailey, H. (2025). Using artificial intelligence for scholarly writing. AJN American Journal of Nursing, 125(11), 52. https://doi.org/10.1097/ajn.0000000000000179

O'Halloran, K. L., Tan, S., & E, M. K. L. (2015). Multimodal analysis for critical thinking. Learning, Media and Technology, 42(2), 147. https://doi.org/10.1080/17439884.2016.1101003

Orbik, Z. (2024). Husserl's concept of transcendental consciousness and the problem of AI consciousness. Phenomenology and the Cognitive Sciences, 23(5), 1151. https://doi.org/10.1007/s11097-024-09993-8

Osbeck, L. M., & Nersessian, N. J. (2017). Epistemic identities in interdisciplinary science. Perspectives on Science, 25(2), 226. https://doi.org/10.1162/posc_a_00242

Page, S. E., & Kallapur, A. (2025). Replace, augment, disrupt: AI & organizational decision-making. Journal of Organization Design. https://doi.org/10.1007/s41469-025-00194-4

Paradice, D. (2010). Emerging systems approaches in information technologies. IGI Global. https://doi.org/10.4018/978-1-60566-976-2

Pasquale, F. (2015). The black box society. Harvard University Press. https://doi.org/10.4159/harvard.9780674736061

Pellegrain, V. (2023). Harnessing the power of multimodal and textual data in industry 4.0 [Doctoral dissertation]. HAL. https://theses.hal.science/tel-04280319

Peters, U. (2022). Explainable AI lacks regulative reasons: Why AI and human decision-making are not equally opaque. AI and Ethics, 3(3), 963. https://doi.org/10.1007/s43681-022-00217-w

Pinker, S. (2010). The cognitive niche: Coevolution of intelligence, sociality, and language. Proceedings of the National Academy of Sciences, 107, 8993. https://doi.org/10.1073/pnas.0914630107

Piredda, G., & Francesco, M. D. (2020). Overcoming the past-endorsement criterion: Toward a transparency-based mark of the mental. Frontiers in Psychology, 11. https://doi.org/10.3389/fpsyg.2020.01278

Plotkin, H. (1995). Darwin machines and the nature of knowledge: Concerning adaptations, instinct and the evolution of intelligence. Penguin Books.

Qamar, T., & Bawany, N. Z. (2023). Understanding the black-box: Towards interpretable and reliable deep learning models. PeerJ Computer Science, 9. https://doi.org/10.7717/peerj-cs.1629

Raab, J. (2018). Aristotle, logic, and QUARC. History and Philosophy of Logic, 39(4), 305. https://doi.org/10.1080/01445340.2018.1467198

Ramos, G., Suh, J., Ghorashi, S., Meek, C., Banks, R., Amershi, S., Fiebrink, R., Smith, A., & Bansal, G. (2019). Emerging perspectives in human-centered machine learning. https://doi.org/10.1145/3290607.3299014

Rao, L., Tian, Y., & Atento, R. G. O. (2025). Adoption and perceived effectiveness of AI in education: Personalization, outcomes, and equity. International Journal of Health & Business Analytics, 1(1). https://doi.org/10.65166/qgq89291

Rao, S. (2025). The impact of artificial intelligence tools on human cognitive abilities: A comprehensive review. INNOVAPATH, 1(10), 7. https://doi.org/10.63501/hsdq5611

Richter, J., & Schaller, R. (2025). AI identity threats and reinforcement in organizations: A theoretical model of professional role identity implications. Proceedings of the Annual Hawaii International Conference on System Sciences. https://doi.org/10.24251/hicss.2025.023

Riley, C., Alrefai, O., Reyes, Y. C., & Hammad, E. (2025). Human-AI interactions: Cognitive, behavioral, and emotional impacts. arXiv. https://doi.org/10.48550/arxiv.2510.17753

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. Trends in Cognitive Sciences, 20(9), 676. https://doi.org/10.1016/j.tics.2016.07.002

Robinson, T. R., & Burns, C. M. (2009). Computer algebra systems and their effect on cognitive load. Electronic Workshops in Computing. https://doi.org/10.14236/ewic/ndm2009.62

Rolin, K. (2017). Scientific community: A moral dimension. Social Epistemology, 31(5), 468. https://doi.org/10.1080/02691728.2017.1346722

Roth, C., & Cointet, J. (2009). Social and semantic coevolution in knowledge networks. Social Networks, 32(1), 16. https://doi.org/10.1016/j.socnet.2009.04.005

Ruja, H., & Popper, K. R. (1973). Objective knowledge: An evolutionary approach. Philosophy and Phenomenological Research, 34(2), 278. https://doi.org/10.2307/2106696

Russell, S., & Norvig, P. (1995). Artificial intelligence: A modern approach. Prentice Hall.

Russon, J. (2016). Aristotle's epistemology. Academia. https://www.academia.edu/38260454/Aristotles_Epistemology

Sargeant, H., Jorgensen, M., Shah, A. H., Weller, A., & Bhatt, U. (2025). Unequal uncertainty: Rethinking algorithmic interventions for mitigating discrimination from AI. arXiv. https://doi.org/10.48550/arxiv.2508.07872

Sarkar, A., Xu, X., Toronto, N., Drosos, I., & Poelitz, C. (2024). When copilot becomes autopilot: Generative AI's critical risk to knowledge work and a critical solution. arXiv. https://doi.org/10.48550/arxiv.2412.15030

Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417. https://doi.org/10.1017/s0140525x00005756

Sims, M. (2021). A continuum of intentionality: Linking the biogenic and anthropogenic approaches to cognition. Biology & Philosophy, 36(6). https://doi.org/10.1007/s10539-021-09827-w

Skitka, L. J., Mosier, K. L., & Burdick, M. D. (1999). Does automation bias decision-making? International Journal of Human-Computer Studies, 51(5), 991. https://doi.org/10.1006/ijhc.1999.0252

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. Mind & Language, 25(4), 359. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Stanovich, K. E. (2017). What intelligence tests miss. Yale University Press. https://doi.org/10.12987/9780300142532

Stiegler, B. (1998). Technics and time, 1: The fault of Epimetheus. Stanford University Press. https://doi.org/10.1515/9780804799362

Stroud, B. (2019). Understanding human knowledge in general. In Routledge eBooks (p. 31). Informa. https://doi.org/10.4324/9780429033261-2

Taylor, H., Fernandes, B., & Wraight, S. (2021). The evolution of complementary cognition: Humans cooperatively adapt and evolve through a system of collective cognitive search. Cambridge Archaeological Journal, 32(1), 61. https://doi.org/10.1017/s0959774321000329

Teng, Q., Liu, Z., Song, Y., Han, K., & Lu, Y. (2022). A survey on the interpretability of deep learning in medical diagnosis. Multimedia Systems, 28(6), 2335. https://doi.org/10.1007/s00530-022-00960-4

Toffoli, S. D. (2024). Proofs for a price: Tomorrow's ultra-rigorous mathematical culture. Bulletin of the American Mathematical Society, 61(3), 395. https://doi.org/10.1090/bull/1823

Torraco, R. J. (2016). Writing integrative literature reviews. Human Resource Development Review, 15(4), 404. https://doi.org/10.1177/1534484316671606

Veldman, W., & Swagerman, D. M. (2018). Correcting the incorrect: An exploratory study into the role of the controller in counteracting financial fake news. Archives of Business Research, 6(10). https://doi.org/10.14738/abr.610.5326

Waefler, T., & Schmid, U. (2021). Explainability is not enough: Requirements for human-AI-partnership in complex socio-technical systems. https://doi.org/10.20378/irb-49775

Webb, M., Fluck, A., Magenheim, J., Malyn-Smith, J., Waters, J., Deschênes, M., & Zagami, J. (2020). Machine learning for human learners: Opportunities, issues, tensions and threats. Educational Technology Research and Development, 69(4), 2109. https://doi.org/10.1007/s11423-020-09858-2

Wheeler, M. (2018). The reappearing tool: Transparency, smart technology, and the extended mind. AI & Society, 34(4), 857. https://doi.org/10.1007/s00146-018-0824-x

Wilburn, H. (2020). An introduction to Western epistemology. Oklahoma State University. https://open.library.okstate.edu/introphilosophy/chapter/an-introduction-to-western-epistemology/

Williams, G. Y., & Lim, S. (2024). Psychology of AI: How AI impacts the way people feel, think, and behave. Current Opinion in Psychology, 58, 101835. https://doi.org/10.1016/j.copsyc.2024.101835

Wilson, P. J., Lorenz, K., & Taylor, R. (1977). Behind the mirror: A search for a natural history of human knowledge. Man, 12, 535. https://doi.org/10.2307/2800563

Yamamoto, K. (2017). The transcendental aesthetic and absolute totality of conditions: The problem of metaphysics in the Critique of Pure Reason and a solution. International Journal of Humanities, Social Sciences and Education, 4(2). https://doi.org/10.20431/2349-0381.0402003

Yazdani, S., Shirvani, A., & Heidarpoor, P. (2021). A model for the taxonomy of research studies: A practical guide to knowledge production and knowledge management. Archives of Pediatric Infectious Diseases, 9(4). https://doi.org/10.5812/pedinfect.112456

Younas, A., & Zeng, Y. (2024). A philosophical inquiry into AI-inclusive epistemology. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4822881

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Algorithmic decision-making and the control problem. Minds and Machines, 29(4), 555. https://doi.org/10.1007/s11023-019-09513-7

Zhu, M., Hussin, S., & Hashim, H. (2025). EFL pre-service teachers' professional identity in the age of AI: An integrative review (2015–2025). Environment and Social Psychology, 10(12). https://doi.org/10.59429/esp.v10i12.4361