Distributed Cognition and Human-AI Delegation in Knowledge Work

Akhil S. G., Case Western Reserve University, sg@case.edu
Kalle Lyytinen, Case Western Reserve University, kalle@case.edu

Introduction

Artificial intelligence (AI) changing the nature of work performed in organizations is no longer news in 2025, McKinsey (2025) is calling this *a cognitive industrial revolution*. The study extends the notion of cognitive systems beyond individual actors, focusing on how human expertise operates within socio-technical systems (Dubova et al., 2022) and in this process develops a grounded understanding of how human-AI delegation manifests at knowledge work (Anthony et al., 2023). Autonomous tools and AI are increasingly performing tasks traditionally assumed by human experts, resulting in an environment where the relationships between action and outcome can be unknowable to users (Zhang et al., 2021). This unpredictability can challenge established knowledge frameworks within organizations, necessitating a reevaluation of cognition and decision-making processes (Berente et al., 2021; Lyytinen et al., 2021; Zhang et al., 2021). Hollan et al. (2000) advanced Distributed Cognition as a framework for understanding how cognitive processes are distributed across individuals, artifacts, and environments within sociotechnical systems, thereby showing how such interactions shape work and learning while aligning with the emergent affordances of digital technologies. With this background, we ask the RQ: How does the application of distributed cognition inform the delegation of tasks between humans and AI in knowledge work?

We examine how augmented intelligence systems that enhance rather than replace human decision-making reshape knowledge work in high-risk, high-complexity professions, using radiology as a primary case. In high-stakes domains like radiology, where diagnostic accuracy directly impacts patient outcomes, AI is redefining how clinicians make decisions. Unlike traditional technologies, AI systems act as collaborative partners, transforming trust, in particular epistemic trust, into a dynamic learning process that is inherent to the co-construction of knowledge between human and machine agents (S. G. & Lyytinen, 2025). This reconceptualization of trust reflects a shift from static, one-directional reliance on technological tools to an interactive, iterative, and reflexive epistemic engagement (S. G. & Lyytinen, 2025). Epistemic trust is a multifaceted concept that pertains to the evaluation and acceptance of knowledge transmitted interpersonally. It involves an individual's assessment of the authenticity, reliability, and applicability of this knowledge to their own circumstances. Even though epistemic trust has been explored little in information systems and technology work literature, the rise of AI in knowledge-intensive domains underscores its need. As AI systems perform tasks traditionally handled by experts, human actors must continuously evaluate the reliability, validity, and applicability of AI-generated knowledge (S. G. & Lyytinen, 2025). Epistemic trust, defined as an individual’s capacity to remain open to knowledge that is personally relevant, generalizable, and trustworthy (Bincoletto et al., 2023; Fonagy et al., 2017, 2019) mirrors interpersonal dynamics of social learning but extends into sociotechnical systems where cognitive processes are distributed across humans, machines, and representational tools. Clinicians and AI systems thus co-evolve through iterative interactions, error resolution, and negotiating trust, trying to establish shared accountability.

Through the application of the distributed cognition framework, we found that epistemic trust influences the human-AI delegation process. We argue that epistemic trust in AI functions as a learning mechanism thus influencing adoption and sustained use of AI. The radiologists calibrate their reliance on AI outputs by engaging with its successes, failures, and contextual limitations, while AI systems try to refine their performance through feedback mechanisms embedded in clinical workflows. With a combination of ethnographic techniques and mapping activities of distributed cognition in a longitudinal study, we aim to provide a deeper understanding of human-AI collaboration. The framework evaluation also helps in knowing how to sustain human agency, by reducing overreliance, and fostering resilient interactions in environments where errors have life-altering consequences. Beyond radiology, this study results can be extended to analogous high-stakes domains where learning requires direct, real-time participation between experts and novices. These include military operations, where split-second decisions depend on hybrid human-machine intelligence; aviation, balancing autopilot reliance with pilot situational awareness; and emergency response, coordinating human judgment with predictive analytics during crises, among others. Learning in these scenarios is also situated learning, and with a synthetic agent, the systems are introduced with varying kinds and levels of cognitive architecture (Lyytinen et al., 2021).

A 15 month long ethnographic study was conducted at Midwestern Hospital Department of Radiology (an ACR Recognized Center for Healthcare-AI) with 250 radiologists, through direct observation of reading rooms (> 50 hours across four specialties), and longer semi structured (>13 hours) and shorter informal interviews (>20 radiologists). The epistemic form of trust is disrupted by errors in AI models and may stabilize or collapse based on error types and clinicians' ability to reconcile discrepancies. These errors differ from traditional technical malfunctions, as they involve cognitive challenges in truth determination, requiring radiologists to second-guess AI outputs and assert their expertise. "I doubt, therefore I think, therefore I am," as René Descartes famously declared, underscores doubt as a catalyst for critical thinking, knowing and learning. In the context of epistemic trust also we see doubt playing a similarly pivotal role: it drives individuals to question, validate, and refine their understanding. Epistemic trust, however, provides a mechanism to manage doubt, enabling learning by balancing skepticism with confidence in the reliability of knowledge claims. Yet, this trust is double-edged. On one hand, it fosters learning by reducing uncertainty; on the other, it risks halting learning if trust becomes over reliant, discouraging critical inquiry. In this system, as AI lacks intentionality, epistemic trust must be dynamically calibrated to ensure sustained use, that is neither too rigid to stifle doubt nor too fragile to undermine collaboration. The delicate balance ensures that doubt remains a productive force, driving learning without derailing progress. Recent studies have shown the potential cognitive costs that come with AI tool reliance (Gerlich, 2025; Kosmyna et al., 2025; Lee et al., 2025).

Knowing is Socially Distributed

In traditional cognitive anthropology, knowing was centered on the question - *What does a person have to know?* with the assumption that the locus of knowledge resided inside the individual. Hutchins (1995b, 1995a) contended that these assumptions prevent our understanding of human cognition. Rather, he highlighted how people come to know what they know and the role of the environment in which knowing takes place. Knowing is seen in practice and is intrinsically a cultural and social process. It involves coordination among various media[1], both internal and external to the individual, bringing these mediating structures into alignment so they can influence and be influenced by one another. The purposes for which people employ their cognitive abilities and the explanations for human cognitive achievement should be explored by examining the real world. Socially distributed cognition refers to the fact that cognition is socially distributed. Hutchins (1995b) argues that systems of socially distributed cognition may have interesting cognitive properties of their own. [detailed explanation removed because of page constraint]

Philosophically, this account resonates with Ludwig Wittgenstein's later work, especially *On Certainty* (1969), which dismantles the notion of private knowledge and emphasizes the social embeddedness of epistemic justification (Wittgenstein, 1969). Both Hutchins and Wittgenstein challenge foundationalist models and foreground the contextual, interactive, and normative dimensions of knowing. While Hutchins does not explicitly cite Wittgenstein, their shared sensibility toward epistemic practices as socially organized and materially mediated suggests a deep conceptual affinity. This convergence between Hutchins and Wittgenstein reveals a deeper epistemological shift: both reject the Cartesian model of cognition as an isolated, individual endeavor and instead position knowing as an emergent property of dynamic interactions, between minds, tools, social norms, and cultural practices. Where Wittgenstein dismantles the philosophical myth of 'private knowledge' by showing how certainty is rooted in shared action - "Giving grounds, however, justifying the evidence, comes to an end; but the end is not certain propositions' striking us immediately as true... it is our acting, which lies at the bottom of the language-game" (On Certainty Pg. 204), Hutchins provides the empirical counterpart. His ethnography of ship navigation demonstrates how 'acting' (e.g., coordinating calculations across crew, charts, and instruments) is the cognitive process - not merely its output. Importantly, both thinkers dissolve the boundary between the internal and external in cognition. Wittgenstein's *world-picture* (Weltbild), the tacit framework of beliefs absorbed through cultural training, finds its operational counterpart in Hutchins' *mediating structures*

[1] media refers to the diverse cognitive and material resources—internal (mental), external (tools, artifacts), social (language, norms), and embodied (skills, practices)—that interact to enable distributed cognition (Hutchins 1995)

(e.g., nautical charts, ritualized protocols, hierarchical roles). These structures are not passive tools but active participants in cognition, just as Wittgenstein's language-games are not mere communication systems but constitutive of meaning itself. When Hutchins describes a novice navigator learning through *culturally prescribed behaviors*, he echoes Wittgenstein's insistence that judgment is taught "not by rules alone, but by being trained in a practice" (OC Pg. 95). The implications are profound: if cognition is distributed, then its failures and triumphs cannot be attributed to individual minds alone. A navigation error, for instance, might stem from a misaligned tool, a breakdown in team coordination, or an ambiguous cultural norm. Similarly, Wittgenstein's observation that "knowledge is in the end based on acknowledgement" (OC Pg. 378) underscores that even the simplest certainty (e.g., this is a hand) relies on a scaffold of social agreement. Hutchins' works thus empirically validate Wittgenstein's philosophical claim: the *real world* of cognition is not a solitary mind confronting nature, but a collective, culturally saturated dance of doing - where knowing is enacted, not stored. In this context of AI and knowledge work, the process of knowing or what is known is foundationally questioned. Knowing here is knowing the unknown stochastically predicted output. [Continued literature review on Extended Mind thesis by (Clark & Chalmers, 1998) and its critique from the likes of (Rupert, 2004) are relevant, but removed due to page constraints, we also explore the role of errors and redundancy in system to tighten the claims made in findings]

Initial Findings

The implementation of AI solutions in radiology reading rooms constitutes not merely an addition to existing practices, but a systemic reconfiguration of cognitive, epistemic, and organizational processes. Drawing on distributed cognition (Hutchins, 1995; Hollan et al., 2000), this section examines how radiologists, AI systems, digital artifacts, and institutional workflows form an interconnected epistemic system that transforms the nature of knowing, trusting, learning, and working in diagnostic practice. The findings can be understood as a process in steps, however not following specific order:

Step (1) Reconfiguring Knowing: Radiological knowledge has historically been constructed through embodied perceptual labor and the sequential coordination of multimodal representations like images, EMR notes, priors, and narrative dictations. AI systems introduce an additional representational layer of triage alerts, heatmaps, and widgets that reconfigures this system of attention. The challenge is profound: radiologists now integrate a new, often unreliable representational medium. As one participant noted, *"The AI flashes at my face, and then I start the scan biased already."* This pre-attentive bias reflects distributed cognition's insight that cognition is shaped by external representations' salience. Yet AI's inconsistencies force radiologists into perpetual meta-cognitive work: *"You get one impressive call and then the exact opposite where it misses a flagrantly positive thing"(e.g., an "undeniable obvious fracture").*

AI's epistemic opacity exacerbates this tension. Unlike traditional tools, AI lacks transparent intentionality that, it cannot explain its reasoning. As a senior radiologist observed, *"You are now coordinating your thinking with an unreliable narrator.*" These transforms knowing from perceptual mastery to representational triage: radiologists must continually assess which representations (human, algorithmic, or institutional) to privilege in a given moment. The challenges with knowing and changes in the knowledge representation has variation across AI solutions. These variations are largely associated with availability of redundant yet accessible alternate representations of a diagnostic result. This not only enhances explainability but also taps into validation of knowledge claims or opportunities for it.

Step (2) Reconfiguring Trust: Trust in radiology is neither static nor unidirectional, it is continuously negotiated through interaction, performance, and institutional validation. AI disrupts traditional trust paradigms by introducing functional trust (human-to-system) alongside relational trust (human-to-human). This functional form of trust is epistemic in nature grounded on reliability, validation of knowledge claims and truthfulness of the result. The volatility of this trust is stark: "*AI is helpful until it is not*" captures its conditional utility. Participants described "*residual skepticism*" born of erratic performance inconsistency where AI oscillates between "*impressive call[s]*" and glaring misses. This necessitates provisional trust: a dynamic calibration where radiologists extend trust contextually, verified through continuous validation.

Epistemic Trust also stratifies by expertise, as it is knowledge based. Juniors face dual calibration: navigating AI while inferring senior expectations without feedback *("It would be nicer if someone was sitting here with us as we were using it in real time").* This mirrors distributed cognition's emphasis on trust as systemic, shaped by socio-technical infrastructure rather than individual disposition. Distributed cognition treats trust as something that arises through coordinated interaction within a system (e.g., radiologists relying on AI tools), rather than just depending on whether an individual inherently trusts technology or not. However, this might have an impact on learning and knowledge acquisition.

Step (3) Reconfiguring Learning: AI in novices has a potential to create an augmented Zone of Proximal Development (ZPD), but one fundamentally distinct from Vygotsky's original formulation (Vygotsky, 1978). Whereas traditional ZPD relies on *a more knowledgeable other* (e.g., a mentor), AI offers non-intentional scaffolding, providing feedback without dialogue. Juniors acknowledged AI's pedagogical potential: *"If you miss something, you can go back and see if the AI flagged it."* Yet this feedback is often disjointed from workflow *("We don't see the feedback in real-time... It's not embedded into our flow")*, undermining its utility. Seniors fear epistemic atrophy: *"Residents need to learn pattern recognition the hard way. If AI tells them what to look for right away, they're skipping the critical step."* This tension between AI as cognitive collaborator and crutch, reflects distributed cognition's warning: learning is not internalization of facts but participatory engagement with scaffolded environments. During observation, we have seen novices developing more dependency on AI solutions, as there are instances where the junior either waits for AI to validate or gets pushed back for trusting the AI's call. In augmented workplaces and communities of practice that prioritize on-the-job learning, addressing the lack of intentionality in AI systems becomes inevitable. The value derived from AI varies between expertise, while novices rely on AI for foundational scaffolding, experts leverage it to refine their decision-making, thus creating a layered learning ecosystem unique to augmentation.

The AI-integrated radiology workplace is a metahuman (Lyytinen et al., 2021) cognitive system - a network where diagnostic insight emerges from human-AI coordination, redundant error-checking, and continual trust negotiation. AI does not replace radiologists; it reconfigures their epistemic practices: as knowing becomes representational triage amid unreliable artifacts, trust shifts from relational to functional, requiring continuous validation and learning operates in an augmented ZPD lacking intentional scaffolding thus leading to new strategies to manage cognitive overload. This system's resilience hinges on integration, not algorithmic accuracy alone, but alignment with human workflows, attentional rhythms, and institutional norms. As one radiologist starkly noted*: "The minute you become complacent, you miss a fracture."* The distributed cognition framework thus reframes AI's role: not as a tool, but as an actor in a reorganized epistemic ecology.

References

Anthony, C., Bechky, B. A., & Fayard, A.-L. (2023). "Collaborating" with AI: Taking a System View to Explore the Future of Work. Organization Science, orsc.2022.1651. https://doi.org/10.1287/orsc.2022.1651

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Special Issue Editor's Comments: Managing Artificial Intelligence. 19. https://doi.org/10.25300/MISQ/2021/16274

Bincoletto, A. F., Zanini, L., Spitoni, G. F., & Lingiardi, V. (2023). Negative and positive ageism in an Italian sample: How ageist beliefs relate to epistemic trust, psychological distress, and well-being. Research in Psychotherapy: Psychopathology, Process and Outcome, 26(2). https://doi.org/10.4081/ripppo.2023.676

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

Dubova, M., Galesic, M., & Goldstone, R. L. (2022). Cognitive Science of Augmented Intelligence. Cognitive Science, 46(12), e13229. https://doi.org/10.1111/cogs.13229

Fonagy, P., Luyten, P., Allison, E., & Campbell, C. (2017). What we have changed our minds about: Part 2. Borderline personality disorder, epistemic trust and the developmental significance of social communication. Borderline Personality Disorder and Emotion Dysregulation, 4(1), 9. https://doi.org/10.1186/s40479-017-0062-8

Fonagy, P., Luyten, P., Allison, E., & Campbell, C. (2019). Mentalizing, Epistemic Trust and the Phenomenology of Psychotherapy. Psychopathology, 52(2), 94–103. https://doi.org/10.1159/000501526

Gerlich, M. (2025). AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. Societies, 15(1), Article 1. https://doi.org/10.3390/soc15010006

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. ACM Transactions on Computer-Human Interaction, 7(2), 174–196. https://doi.org/10.1145/353485.353487

Hutchins, E. (1995a). Cognition in the wild. MIT Press.

Hutchins, E. (1995b). How a Cockpit Remembers Its Speeds. Cognitive Science, 19(3), 265–288. https://doi.org/10.1207/s15516709cog1903_1

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task (No. arXiv:2506.08872). arXiv. https://doi.org/10.48550/arXiv.2506.08872

Lee, H.-P. (Hank), Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 1–22. https://doi.org/10.1145/3706598.3713778

Lyytinen, K., Nickerson, J. V., & King, J. L. (2021). Metahuman systems = humans + machines that learn. Journal of Information Technology, 36(4), 427–445. https://doi.org/10.1177/0268396220915917

McKinsey. (2025). AI in the workplace: A report for 2025 | McKinsey. McKinsey. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/superagency-in-the-workplace-empowering-people-to-unlock-ais-full-potential-at-work

Rupert, R. D. (2004). Challenges to the Hypothesis of Extended Cognition. The Journal of Philosophy, 101(8), 389–428. https://doi.org/10.5840/jphil2004101826

S. G., A., & Lyytinen, K. (2025). Continued Augmentation in Metahuman Systems: Phase of Trust Building. Academy of Management Proceedings.

Vygotsky, L. S. (1978). Mind in Society: The Development of Higher Psychological Processes (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.; Revised ed. edition). Harvard University Press.

Wittgenstein, L., Anscombe, G. E. M., & Wright, G. H. von. (1969). On certainty. Blackwell.

Zhang, Z., Yoo, Y., Lyytinen, K., & Lindberg, A. (2021). The Unknowability of Autonomous Tools and the Liminal Experience of Their Use. Information Systems Research, 32(4), 1192–1213. https://doi.org/10.1287/isre.2021.1022