

# **Production of Knowledge in Human-Machine Collaborations: The Hybrid Intelligent Team as Epistemological Form**

## **Abstract**

Whether human-machine collaborations can produce genuine knowledge, knowledge with the same epistemic validity as that produced without intelligent machines, is one of the defining contested questions of this moment in the history of ideas. This paper argues that the answer depends not on the capabilities of individual AI systems but on the organizational conditions under which human and machine intelligence collaborate. We develop the theory of the Hybrid Intelligent Team (HIT)—a structured multi-agent configuration characterized by differentiated functional roles, asymmetric memory architectures, and a human coherence anchor—as an epistemological form that creates conditions under which warranted knowledge claims can be produced through human-machine collaboration.

The paper makes three contributions. First, it identifies the structural gap in existing literature across three research communities—authorship and attribution, capability assessment, and process and workflow—none of which has theorized the organizational conditions for epistemic warrant in human-machine collaboration. Second, it develops the theoretical framework, arguing that the HIT's three structural conditions are not independent virtues but foundations for a fourth and more radical condition: loop-level ideation, in which knowledge is produced not by the human or the machine but by the iterative cycle connecting them. Third, it grounds the theory in four empirical cases—the "Can Machines Have Ideas?" Substack series, the Anti-ESG Ecosystem book, Operation Epic Fury, and a Pure Mathematics Experiment—that vary systematically on the dimensions the theory identifies as consequential.

The paper also generates eight testable propositions for the empirical agenda the theory motivates, including a proposition on the transferability of the orchestrator role across different expertise profiles and cognitive styles, the foundational empirical question the theory cannot answer from a single case. It also identifies four open questions that define the productive frontier of this research program: the Recognizer Problem, the tacit knowledge problem, the institutional mismatch between HIT production and discipline-organized evaluation, and the time dimension of a developing human-machine collaboration system. The paper was itself produced under the organizational conditions it describes, and its methodological appendix documents that production process as evidence bearing on the theory's central claims. The complete production record, including adversarial review, independent reader evaluation, and the changes made in response to it, is documented in Appendices A through D.

Keywords: knowledge production, Hybrid Intelligent Teams, human-AI collaboration, social epistemology, warranted belief, distributed cognition, Narrative AI Ethnography, loop-level ideation, Recognizer Problem

JEL Classification: O30, O31, O32, O33, O39

Word Count: approximately 18,000 words (excluding abstract, references, and appendices)

## I. Introduction

Whether human-machine collaborations can produce genuine knowledge—knowledge with the same epistemic validity as that produced without intelligent machines—is one of the defining contested questions of this moment in the history of ideas. The question is not merely academic. It bears on how universities evaluate scholarship, how journals assess submissions, how courts weigh expert testimony, how policymakers treat evidence, and how all of us decide what to believe. The institutions that have historically organized knowledge production were designed for a world in which the relevant agents were human. That world is ending, and the institutions have not caught up.

The evidence of institutional strain is already visible and accumulating. AI assistance has enabled a dramatic increase in the volume of manuscripts entering the academic pipeline, while disclosure norms—however well-intentioned—cannot be reliably enforced and are inconsistently observed. The result is a system under asymmetric pressure. Submissions are AI-assisted at scale, but review remains a human obligation conducted under guidelines that prohibit reviewers from using the same tools. A senior editor at a leading management journal may receive upward of 350 submissions per year, each requiring substantive evaluation, with no legitimate recourse to the tools that produced them. This is not a temporary adjustment problem. It is a structural consequence of asking the wrong institutional question, which is the same as asking the wrong intellectual question. The debate about AI and knowledge production has focused on what AI systems can generate. The institutions have responded by trying to manage the volume and label the provenance. Neither the debate nor the institutional response has asked what organizational conditions would make human-machine collaboration epistemically legitimate in the first place. Until that question is answered, the institutional fixes will remain patches on a cracking foundation.

The debate about what replaces existing arrangements has proceeded largely by asking what AI systems can generate. Can they produce original arguments? Can they pass peer review? Can they make claims that extend beyond the existing knowledge frontier rather than recombining what is already known? These are important empirical questions, and the answers accumulating are striking. AI systems can do more than most observers anticipated a decade ago, and the boundary between retrieval and genuine generation remains genuinely contested. But we argue that this is the wrong question, or at least an insufficient one. The capacity of an individual AI system to generate a particular kind of output tells us relatively little about whether the conditions exist for that output to constitute knowledge. Knowledge is not merely a property of texts. It is a property of the relationship between texts and the epistemic processes that produce and warrant them.

The right question is what organizational conditions make it possible for human and machine intelligence to function as an epistemologically coherent unit—one capable not merely of producing fluent, coherent, performance-effective outputs, but of producing genuinely warranted knowledge claims. That question has not been the literature's question, and the gap is structural rather than incidental. The three research communities that have engaged most seriously with AI and knowledge production—those concerned with authorship and attribution, with capability assessment, and with process and workflow—each illuminate part of the problem and each stops short of the epistemological question for reasons internal to its own framing. We examine that literature in Section II. The diagnosis matters because the gap it reveals is precise: not a general

absence of epistemology for AI-assisted research, but the specific absence of a theory of organizational form as the condition of epistemic warrant.

We argue that the Hybrid Intelligent Team (HIT)—a structured multi-agent configuration in which distinct AI systems occupy differentiated functional roles, memory architectures vary systematically across agents, and a human orchestrator serves as coherence anchor and final recognizer—creates epistemological conditions for which no existing theory of human-AI collaboration provides an adequate account. The three structural conditions are not independent virtues of good organizational design. They are foundations for a fourth and more radical condition: loop-level ideation, in which knowledge is produced not by the human or the machine but by the iterative cycle connecting them. Unlike ordinary AI-assisted work, in which a careful reader could still separate what the human contributed from what the machine contributed, loop-level ideation produces knowledge claims that are irreducibly joint; the ideas emerge from the interaction itself, and decomposing the output by contributor loses what is analytically significant about it. When the loop is functioning, the question of what the human contributed versus what the machine contributed does not merely become difficult to answer. It becomes the wrong question to ask. The unit of knowledge production is the loop, and no decomposition of individual contributions can adequately describe what the loop produces.

This theoretical claim is grounded in a documented body of empirical work. We draw on four cases produced using exactly the organizational form we theorize: the “Can Machines Have Ideas?” series, a seven-piece body of collaborative inquiry conducted on Substack in which the human orchestrator was progressively removed from the interpretive loop; a book and case study on the anti-ESG ecosystem, produced under conditions of deep domain expertise with the most rigorously documented production process in our corpus; Operation Epic Fury, a four-piece series produced under conditions of thin domain expertise that functions as a natural experiment on what the coherence anchoring function requires and what it can substitute for; and A Pure Mathematics Experiment, which examines the coherence anchoring function and the Recognizer Problem under a third distinct condition—atrophied domain expertise that partially reactivates under loop pressure, producing a published mathematical note whose epistemic status is itself an open question the experiment was designed to surface. Taken together, the four cases provide empirical purchase on the conditions under which the loop produces knowledge-like outputs, the conditions under which warrant can be claimed for those outputs, and the structural tension—which we call the Recognizer Problem—that the theory cannot yet resolve.

The paper’s structure is as follows. Section II characterizes the three camps of existing literature and states the gap precisely. Section III develops the theoretical foundations: what distinguishes knowledge from sophisticated text, what distinguishes new knowledge from recombination, and the Recognizer Problem as the field’s open question. Section IV develops the theory of the HIT as epistemological form. Section V presents the four empirical cases. Section VI draws implications for academic knowledge production, institutional design, and future empirical inquiry. Section VII concludes. A methodological appendix documents how this paper was produced — which is not a supplementary matter but an integral part of the argument.

That self-referential structure deserves a word at the outset. This paper is produced by a Hybrid Intelligent Team consisting of a human researcher and a configured cluster of AI systems occupying differentiated functional roles. Its production process is simultaneously its object of study and an instance of its central claim. When the content of a human-machine collaboration is

itself the study of human-machine collaboration, the compound return is constitutive rather than incidental. Improving as a collaborator and producing knowledge about collaboration are the same activity observed from two angles, a structure that does not obtain when the instrument and the object of inquiry are distinct. That is not a rhetorical flourish. If the theory is correct, the conditions under which this paper was produced are exactly the conditions under which warranted knowledge claims can be produced in human-machine collaboration. If the theory is wrong, or incomplete, the production process is the first place to look for evidence of that. We have tried to make it possible to look.

## **II. What the Existing Literature Gets Wrong: Three Camps**

The question of what happens epistemically when human and machine intelligence collaborate has attracted sustained attention across several research communities. That attention has been productive. It has generated normative frameworks for attribution, substantial empirical evidence about AI capability, and increasingly sophisticated models of human-AI workflow. What it has not generated is a theory of the organizational conditions under which structured human-machine collaboration produces genuinely warranted knowledge claims. Understanding why requires tracing the structural limitation each camp shares—not as a failure of execution but as a consequence of the question each camp is organized to answer.

### **A. Authorship and Attribution**

The authorship and attribution literature was the first to respond systematically to AI's entry into knowledge production, and its response was swift and largely normative. The Committee on Publication Ethics and the International Committee of Medical Journal Editors established, within the span of a year, the canonical position that now governs most academic publishing. AI systems cannot be authors because authorship entails accountability, and accountability requires the capacity for consent, responsibility, and legal standing that no AI system currently possesses (COPE 2023; ICMJE 2024). Disclosure requirements followed as a practical corollary. If AI was used, that use must be declared, even if the AI cannot appear on the byline.

These frameworks addressed a genuine and urgent problem. Academic publishing faced the immediate prospect of AI-generated text entering the peer-reviewed record without any signal to readers, editors, or the communities whose epistemic practices depend on knowing where claims come from and who stands behind them. The attribution camp's response was appropriate to that problem.

Its limitation is that it is a response to a different problem than the one this paper addresses. The attribution question presupposes that knowledge has already been produced and asks who should receive credit for it. It is a normative question about recognition, accountability, and the social organization of credit. It is not an epistemic question about production. COPE and ICMJE are silent on whether the knowledge claims in AI-assisted work are warranted, because that is not what they are trying to determine. The result is a body of guidance that is indispensable for publication practice and irrelevant to epistemology. Whether a paper discloses AI assistance tells us nothing about whether its central claims are genuinely warranted or merely fluent.

## B. Capability Assessment

The capability assessment literature asks a different question: What can AI systems do? This is an empirical rather than normative inquiry, and it has produced striking results. A useful analytical distinction for locating current systems is between AI that is *world-taking* — retrieving and recombining existing representations — and AI that is genuinely *world-making* in its capacity to generate claims that extend beyond the existing knowledge frontier. The empirical evidence about where current systems fall along that spectrum is accumulating rapidly. The AI Scientist project, developed by Sakana AI and reported in Nature’s coverage of autonomous research systems (Lu et al. 2024), demonstrated that AI systems can produce manuscripts that pass peer review—a result that sharpens the world-taking/world-making question considerably. Passing peer review while remaining world-taking would be significant evidence that gatekeeping mechanisms cannot distinguish warranted from unwarranted claims.

More recently, the capability assessment literature has begun to register that organizational structure may be an independent variable in AI performance. Work on OrgAgent (Zhang et al. 2026) treats organizational structure as a condition that modulates multi-agent reasoning outcomes (arXiv:2604.01020). Recent multi-agent research also finds that self-organizing configurations can outperform rigid hierarchies on complex tasks, a finding that creates an immediate tension with any theory that assigns fixed functional roles to agents in a multi-agent system, and one we engage directly in Section IV.

The structural limitation of the capability assessment camp is that its question remains at the level of the agent, or at most the system, rather than at the level of epistemological form. Even when organizational structure enters the analysis, the dependent variable is performance—accuracy, task completion, benchmark scores—not epistemic warrant. Demonstrating that an AI system can produce a manuscript that passes peer review, or that a self-organizing multi-agent configuration outperforms a hierarchical one on a reasoning task, does not answer the question of whether the claims produced are warranted in the sense required for knowledge. Performance and warrant are related but not identical. A system can be reliably accurate in ways that fall short of knowledge, and the conditions under which reliable accuracy becomes warranted belief have not been the capability camp’s question to answer.

## C. Process and Workflow

The process and workflow literature is the camp most directly concerned with human-machine collaboration as an organizational phenomenon, and in several respects it comes closest to the territory this paper occupies. Most closely, it shares a neighborhood with the distributed cognition tradition in cognitive science, whose relationship to this paper’s argument requires separate treatment below. Gonzalez et al. (2026), writing in *PNAS Nexus*, develop a complementarity framework that identifies organizational conditions (role differentiation, structured interaction, and human-AI pairing) under which human-AI teams produce reliable outputs. Their focus is reliability in decision-relevant contexts, and their contribution is to demonstrate that team-level organization is a meaningful unit of analysis for understanding what human-AI collaboration produces.

Costa et al. (2025) develop AbductivAI, an epistemically-inflected qualitative research workflow in which AI assistance is integrated into abductive inference; their work is the closest empirical analog to our framing and the most self-aware about the epistemic dimensions of human-AI

collaboration among the process camp's contributions. Weidener et al. (2026) document the role of persistent world state in Deep Research systems—as systems engineering treatment of what this paper analyzes as memory asymmetry—and identify it as a structural condition affecting output quality.

Our own prior work belongs in this camp. Eccles (2025a) introduces the Hybrid Intelligent Team as an organizational framework and documents its use across several production contexts, establishing the HIT as a tractable unit of analysis in organizational behavior terms (SSRN working paper, <https://doi.org/10.2139/ssrn.5792345>). Eccles (2025b) develops the Narrative AI Ethnography methodology that grounds the empirical approach across that corpus—treating sustained human-AI collaboration as a field site, with the behavioral and interpretive outputs of the loop as primary data rather than means to an end. That body of work's argument is where this paper's argument begins.

#### **D. The Gap**

What the three camps share, despite their different questions and methods, is a common unit of analysis: the individual agent or the simple dyad. The attribution camp asks about individual AI systems and what credit they are owed. The capability camp asks what individual systems or loosely coupled multi-agent configurations can do. The process camp asks how human-AI dyads or teams can be structured to work effectively. Even the most organizationally sophisticated contributions—OrgAgent, Gonzalez et al., our own prior HIT framework--treat organizational structure as a condition affecting performance rather than as an epistemological form in its own right.

Before stating the gap precisely, the paper must address a prior tradition that a sophisticated reader will already have in mind: the several decades of work in Science and Technology Studies and distributed cognition that have argued, on different grounds, that knowledge is always an emergent property of loops and networks involving non-human actors.

Edwin Hutchins, in *Cognition in the Wild* (1995), demonstrated that the knowledge required to navigate a naval vessel resides not in any individual crew member but in the organization of their labor—the ship's navigation system as a whole is the cognitive unit. Bruno Latour's Actor-Network Theory (ANT) extended this logic further, insisting that non-human “actants”—instruments, documents, laboratory equipment—are genuine participants in the production of scientific knowledge, not merely passive tools (*Science in Action*, 1987; *Pandora's Hope*, 1999). Andrew Pickering's *The Mangle of Practice* (1995) described the “tuning” of human and material agency in scientific work as a dialectical process in which neither the human nor the material term is stable; each shapes and is shaped by the other in the course of inquiry. Karin Knorr Cetina's *Epistemic Cultures* (1999) extended this analysis further still, showing that different scientific communities construct knowledge through organizationally distinct epistemic practices, that the form of knowledge production is not incidental to the knowledge produced but constitutive of it.

These are serious frameworks, and the HIT concept is in genuine dialogue with them. The claim that knowledge is produced by loops and networks rather than individual minds, that non-human actors have genuine roles in epistemic activity, and that the relevant unit of analysis is the organized system rather than the isolated agent are commitments the HIT framework shares with the STS tradition, not departures from it.

What the HIT concept adds, and what that tradition does not supply, is specific to the nature of the non-human agents now in the loop. Hutchins' navigation team includes instruments—sextants, charts, and alidades—that are precisely constrained in what they can do. They are, in Latour's term, mute mediators, their agency exhausted by their physical and mechanical properties. Latour's laboratory actants speak only through the experimental apparatus; Pickering's material agency resists and accommodates but does not propose. The AI agents in a HIT are different in kind. They are fluent, generative, and capable of adversarial challenge; they possess differentiated epistemic orientations—one agent serves as a principal collaborating writer and synthesizer, another as an adversarial challenger, another as an independent first reader—and they operate under asymmetric memory architectures that make each agent's epistemic relationship to the project structurally distinct.

A distributed cognition framework that was built to describe the cognitive properties of organizations of humans and instruments does not automatically extend to organizations whose non-human members can formulate objections, propose reframings, and sustain extended analytical dialogue. The HIT is, in this sense, the first digital-native instance of the distributed cognition insight, but it requires a theory of its own because the character of its non-human participants is discontinuous with anything the STS tradition had available to theorize.

The consequence is a gap that Koskinen (2024), writing in *Social Epistemology*, identifies from within the discipline best positioned to notice it. We have no satisfactory social epistemology of AI-based science. Koskinen's diagnosis is that the existing epistemic frameworks—testimony, distributed cognition, and extended mind—were developed for human-only epistemic communities and do not transfer cleanly to configurations in which AI systems play active, non-trivial roles in knowledge production. The gap he identifies is structural, not merely a matter of insufficient empirical work. The STS tradition, for all its insight, does not resolve it precisely because that tradition's non-human actants were never fluent enough to generate the Recognizer Problem.

This paper responds to that gap with a specific theoretical claim. The gap is not simply that we lack epistemology for AI-assisted research in general. The gap is more precise:

*No existing literature contains a theory of the organizational conditions—specifically, differentiated roles, asymmetric memory architecture, and a human coherence anchor—under which structured human-machine collaborations produce genuinely warranted knowledge claims, as distinct from outputs that are fluent, coherent, performance-effective, and indistinguishable by existing gatekeeping mechanisms from knowledge. Nor does any existing literature recognize that these three conditions are not endpoints but foundations. When they operate together, they make possible a fourth and more radical condition, loop-level ideation, in which knowledge is produced not by the human or the machine but by the iterative cycle connecting them. At that level of organization, the question of what the human contributed versus what the machine contributed does not merely become difficult to answer. It becomes the wrong question to ask. It is this reframing of the unit of knowledge production, from agent to loop, that the existing literature has not theorized and that this paper attempts.*

### III. Theoretical Foundations: What Knowledge Requires

#### A. Knowledge versus Sophisticated Text

The question of whether human-machine collaborations produce genuine knowledge cannot be answered without first specifying what knowledge requires. This is not a terminological nicety. It is the condition of possibility for the entire inquiry, because the most significant challenge posed by AI involvement in knowledge production is not that AI systems produce wrong outputs, it is that they produce outputs that are fluent, coherent, internally consistent, and indistinguishable by existing evaluation mechanisms from genuine knowledge, while potentially lacking the properties that make knowledge rather than sophisticated text.

The classical account of knowledge as justified true belief provides a starting point. A knowledge claim is warranted when it is justified, when there are good reasons, traceable and interrogatable, for believing it, and when that justification is robust enough to survive adversarial scrutiny from people who know the domain and are trying to find holes. Reliability is a further condition. A knowledge claim is not merely a lucky hit but the output of a process that would tend to produce true beliefs under relevantly similar conditions (Goldman 1999). Reproducibility matters for the same reason. If the conditions that generated the claim cannot be reconstructed or approximated, the claim's epistemic credentials are weakened regardless of its surface plausibility. And uptake—whether the claim changes what others do, what questions they ask, what relationships they see—is the community-level test of whether something has entered the shared epistemic commons rather than merely circulated within it (Kitcher 1993; Longino 2002).

These conditions operate at two distinct levels and keeping that distinction precise matters for everything the paper argues. The first three (justification, reliability, and reproducibility) are process-level conditions; they can be assessed from within the production system, by examining the process through which a claim was generated and evaluated. The fourth, uptake, is community-level. It is the external test through which warrant is conferred rather than merely claimed, and it cannot be performed by any participant in the production process, however expert. A claim that meets the process-level conditions is *warrantable*; it is structurally positioned for warranted status. Whether it achieves that status depends on the community of inquiry. This distinction between warranted and warrantable will carry weight throughout the paper, particularly in the analysis of the human coherence anchor's recognition function in Section IV.C.

These warrant conditions are not exotic philosophical requirements. They are what peer review, replication, citation, and adversarial commentary are designed to test. The problem is that existing gatekeeping mechanisms test for proxies of these conditions rather than the conditions themselves. Peer review evaluates whether an argument is coherent and well-structured, whether it engages the relevant literature, whether its methods are appropriate, and whether its conclusions follow from its evidence. These are necessary but not sufficient conditions for warranted knowledge. A system capable of producing fluent, well-structured, literature-engaged, methodologically appropriate text can pass peer review without its outputs being warranted in the deeper sense—without the justification being traceable, the reliability being grounded in a process that tends to produce true beliefs, or the claim being the kind of thing others can build on rather than merely cite.

This is precisely what Koskinen (2024) identifies when she argues that we have no satisfactory social epistemology of AI-based science. The established framework for collaborative science

holds that knowledge production in large teams requires relationships of trust, not merely reliance, and that trust is only possible between agents who can be held accountable. Epistemically opaque AI systems are not appropriate candidates for trust in this sense, yet scientists are increasingly epistemically dependent on them. The problem is not that AI outputs are unreliable in the narrow sense of being factually wrong more often than human outputs. The problem is structural. The conditions under which AI outputs are generated do not straightforwardly ground the warrant conditions that make a claim genuine knowledge rather than sophisticated text.

This structural problem has a name in the recent literature. The concept of Epistemia, which describes the condition in which linguistic plausibility substitutes for epistemic evaluation, producing the feeling of knowing without the labor of judgment, captures what happens when the evaluation function is delegated to a system optimized for plausible output rather than warranted output. A system subject to Epistemia can produce confident, well-structured, internally coherent claims that lack the grounding in evidence, the traceability of justification, and the reliability of process that genuine knowledge requires. The challenge for any theory of AI-mediated knowledge production is to specify the organizational conditions that compensate for this structural deficit, the conditions under which human judgment re-enters the loop at the points where the process-level warrant conditions are assessed, rather than merely simulated by fluent output.

## **B. New Knowledge versus Recombination**

If the first distinction, knowledge versus sophisticated text, concerns the warrant conditions for any knowledge claim, the second concerns what makes a knowledge claim genuinely new rather than incrementally additive. This distinction matters because a great deal of what circulates as new knowledge in academic publishing is, on examination, sophisticated recombination—existing concepts combined in novel ways, existing evidence organized under a slightly different frame, or existing arguments extended to new domains. These are not without value. But they are not new knowledge in the strong sense, the sense in which a claim changes what questions are askable, what relationships are visible, or what problems become tractable.

The empirical work documented in the “Can Machines Have Ideas?” series generated a four-level hierarchy that provides a useful operational framework for this distinction (Eccles 2025–26). At the base is synthesis: aggregating and organizing existing knowledge without genuine combination. Most AI output defaults to this level, and so does a great deal of mediocre academic work. Above it is recombination: familiar elements in genuinely new configurations, producing outputs that were not predictable from the inputs even if all the ingredients were already available. Above that is reframing: the same facts organized under a different explanatory structure, one that makes something previously murky suddenly tractable. This is where most genuinely good scholarly work lives—the phenomenon does not change, but what changes is the framework through which it is seen, and that change has consequences for what can be done with it. At the apex is paradigm shift: the rarest category, in which the claim restructures what questions are even askable. Kuhnian in the full sense.

This hierarchy is not merely descriptive. It is diagnostic. It allows us to ask, of any given output, not just whether it is warranted but at what level of genuine novelty it operates. The question for a theory of HIT-based knowledge production is not whether HITs can produce synthesis. They clearly can, and so can a single well-prompted LLM. The question is whether HITs can produce reframings with generative consequences, and under what conditions. The claim of this paper is

that they can, that the organizational conditions specified in Section IV are the conditions under which they do so, and that the empirical cases documented in Section V provide evidence for this claim.

The distinction between reframing and recombination is the operative threshold for what we are calling new knowledge. A reframing is not merely a new combination of existing elements. It is a reorganization of existing elements under a structure that makes previously intractable questions tractable. It creates new discriminations, opens new lines of inquiry, and changes what the next question is. The test is generative consequence. Does the reframing produce outputs downstream of itself that would not have been produced without it? Does it move from language to leverage?

This test is auditable in principle. It is not a matter of subjective impression but of documented consequence. The “Can Machines Have Ideas?” series provides a running instance of this test. Each piece changed what the next question was, and the changes were not predictable from any single prior piece. The loop was generative in the precise sense; it produced downstream movement that was not contained in any of its individual components.

### **C. The Recognizer Problem**

The theoretical framework developed in this paper rests on a tension it does not resolve and does not claim to resolve. We call it the Recognizer Problem, and it is the most important unresolved question in the theory of HIT-based knowledge production. The closely related question of when AI systems can be credited as proper bearers of scientific discovery, as distinct from instruments through which human discoverers work, is developed in Duede and Friedman (2025), and it bears directly on the Recognizer Problem. If the human cannot evaluate what the loop has produced, the question of discovery credit cannot be resolved by inspecting the output alone.

The loop-level theory of ideation holds that the right unit of analysis for whether a HIT produces new knowledge is not any individual agent, human or machine, but the loop itself—the iterative cycle of generation, challenge, verification, and recognition through which outputs are produced, evaluated, and either discarded or incorporated into the accumulating epistemic structure of the project. In this framework, the author of a genuinely new idea is often the loop rather than any individual node. The human supplies aims and judgment; the AI systems supply generative pressure, adversarial challenge, and surprising continuations; the world supplies constraint; and the knowledge claim emerges as the loop’s stable output, the thing that survived contact with all of these forces and remained standing.

This is a compelling account of how knowledge production actually works in practice. But it depends on a condition that the theory cannot fully discharge. It depends on the human recognizer. The loop cannot position its own outputs for epistemic warrant without human recognition. Recognition, the judgment that a given output meets the process-level conditions for warranted knowledge and is ready for community evaluation, is irreducibly a human function in the current state of AI development. And this creates a structural tension. If the loop can only produce knowledge that the recognizer is already capable of recognizing as knowledge, then the loop is bounded by the recognizer’s prior framework. It can surface things the recognizer had not yet articulated, and it can force articulation of tacit knowledge by creating the need to evaluate and respond. But if it cannot produce knowledge that genuinely exceeds the recognizer’s prior framework, the recognizer could not have produced alone and could not have recognized as such

without the loop, then the HIT is a powerful amplification device but not a genuinely generative epistemological form.

Gemini Pro's "Ghost in the Prompt" inversion is the sharpest version of this problem. The suggestion that the AI system may be training the human to provide better context, effectively turning the human into a high-fidelity data-retrieval agent for the AI's generative engine, is not merely a clever provocation. It is a genuine question about the direction of epistemic dependence in the loop. Who is recognizing what, and whose framework is expanding?

We treat this as a structural tension in the theory rather than a defect to be papered over. The Recognizer Problem does not undermine the paper's central claim that the HIT creates epistemological conditions that do not exist in simpler configurations. It qualifies it. Those conditions make possible kinds of knowledge production that are unavailable to single agents or simple dyads. Whether they make possible knowledge production that genuinely transcends the recognizer's prior framework, knowledge the recognizer could not have produced or recognized alone, is the field's next empirical question. The empirical cases in Section V provide evidence relevant to this question but do not resolve it. Resolving it will require comparative studies across recognizers with different domain expertise profiles, different methodological orientations, and different relationships to the material which is precisely what the testable propositions in Section VI are designed to enable.

What the theory does claim, and what the evidence supports, is that the HIT creates conditions under which the Recognizer Problem is at its most productive rather than its most constraining. The combination of differentiated roles, asymmetric memory, and human coherence anchoring produces a configuration in which the recognizer is regularly confronted with outputs that exceed what the recognizer would have produced alone, that force articulation of tacit knowledge, and that create discriminations the recognizer could not have made without the generative pressure of the loop. Whether those outputs constitute new knowledge in the strongest sense, knowledge that transcends the recognizer's prior framework rather than merely surfacing and articulating what was latent within it, is a question the theory holds open as its most important unresolved problem.

What the loop demonstrably does, and what neither solo human work nor simple human-AI dyads can achieve, is force articulation of what was tacit and administer adversarial pressure the recognizer cannot self-apply. The outputs the recognizer evaluates are not outputs the recognizer could have produced. Whether evaluating them constitutes transcendence of the recognizer's prior framework or merely its extension is the empirical question the propositions in Section VI are designed to test. But the loop's function is to move the recognizer, not merely to serve them, and that movement, however partial, is the mechanism through which the Recognizer Problem becomes productive rather than merely constraining. A recognizer who has been moved by the loop is a different epistemic agent than one who has not, and the knowledge claims they can position for community evaluation are correspondingly different.

## **IV. The Hybrid Intelligent Team as Epistemological Form**

### **A. Differentiated Roles as Epistemological Variable**

The organizational design literature has established that structure shapes performance. Hierarchical multi-agent configurations outperform flat ones on complex tasks, role differentiation enables division of cognitive labor, and the allocation of distinct functions across agents produces outputs that no single agent could generate alone. These are performance claims. What has not been established, and what this paper argues, is that role differentiation is also an epistemological variable. It does not merely improve the efficiency of knowledge production. It changes what kinds of knowledge can be produced and under what conditions outputs acquire warrant rather than merely plausibility.

The epistemological significance of role differentiation follows from the warrant conditions established in Section III. The process-level conditions for warranted knowledge—justification that is traceable and interrogatable, reliability grounded in a process that tends to produce true beliefs, and reproducibility that allows the conditions of production to be reconstructed and examined—are each more robustly met when distinct functions in the production process are assigned to distinct agents with different epistemic orientations than when a single agent attempts to perform all functions simultaneously.

Within this differentiated structure, one role warrants particular attention: the Principal Collaborating Agent (PCA). The PCA is the agent assigned the coordination and continuity function, the system through which the human orchestrator directs the overall production process, synthesizes contributions from other agents, and maintains coherence across the project. In the Anti-ESG book, Dorothy (an identity-specific agent of ChatGPT) occupied this role; in the production of this paper, Claude does. The PCA designation matters epistemologically because it identifies the agent whose outputs are most directly shaped by, and most directly shape, the human coherence anchor's direction. That structural position makes the PCA's role different in kind from that of the adversarial agent, the verification agent, or the independent reader—not superior in epistemic value, but different in function and in its relationship to the warrant conditions the HIT creates.

Consider the adversarial role. A single agent asked to generate and critique its own output faces a structural conflict. The generative function is optimized for coherence and plausibility, while the critical function requires precisely the disposition to find incoherence and implausibility. When these functions are assigned to the same agent, the critical function is systematically underpowered; the agent's investment in its own output shapes what it is willing to see as a weakness. When they are assigned to distinct agents, one optimized for generative expansion, another for adversarial challenge, the critical function is structurally insulated from the generative function's influence. The output that survives this separation has been exposed to genuine adversarial pressure rather than self-administered critique, and its warrant is correspondingly stronger. One further design principle follows; the epistemological function of adversarial challenge depends on where in the production cycle it is administered. Challenge at the formation stage changes the argument itself; challenge at the mature stage changes the documentation.

The same logic applies to the verification function. Independent verification of factual claims and literature grounding by an agent with no investment in the argument being advanced, and with no knowledge of the argument's conclusions before performing verification, is epistemically different in kind from verification performed by the agent that generated the argument. The former produces a genuine independence condition; the latter produces a form of confirmation that is structurally biased toward the argument it is checking.

The independent reader function completes the picture. An agent encountering the output without accumulated context—without knowledge of what the argument was supposed to be, what decisions shaped its development, or what the producing agents intended—tests whether the argument communicates what it claims to communicate to a reader who has not been party to its production. This is the closest available approximation to the epistemic position of the eventual reader, and it is structurally unavailable to any agent that participated in the production process.

The claim that different AI systems bring structurally distinct epistemic orientations, rather than merely stylistic variations on the same generative capacity, has empirical support from earlier in this research program. Eccles, Dorothy, and Claude (2025b) documented systematic task-type boundaries in AI collaborative behavior. When five systems were asked to engage with identical organizational strategy questions, one system consistently and completely avoided the strategic and organizational dimensions across multiple prompting attempts and across both English and French, while performing well on analytical and editorial tasks in the same exchange. The boundary held even when the system was provided explicit scaffolding showing what organizational engagement looked like from other systems. This behavioral consistency across linguistic contexts and prompting strategies suggests the differentiation is architectural rather than contextual, a finding that the HIT framework's differentiated-roles condition formalizes theoretically. Earlier work in the same research program also documented that Claude, Perplexity, Gemini Pro, and ChatGPT exhibit distinct collaborative personalities stable across interactions, differences that pre-theoretical observation identified and that differentiated role assignment is designed to operationalize (Eccles, Dorothy, and Claude 2025a).

The recent literature on self-organizing multi-agent systems raises an important complication here. Recent work finds that hybrid protocols enabling agents to develop their own role orientations dynamically can outperform both rigid pre-specified hierarchies and unconstrained autonomy on complex tasks, with the degree of emergent autonomy scaling with model capability (Zhang et al. 2026; Eccles 2025a). This finding does not undermine the case for differentiated roles. It qualifies the mechanism. The HIT is not a rigid hierarchy with fixed role assignments. It is a configuration in which minimal scaffolding (the human orchestrator's assignment of broad functional orientations rather than detailed behavioral specifications) creates conditions for the emergence of epistemically productive role differentiation, with the human coherence anchor providing precisely that scaffolding by establishing overall epistemic direction and evaluative criteria without pre-specifying how each agent should contribute to meeting them.

## **B. Memory Asymmetry as Epistemological Variable**

The Hybrid Intelligent Team as described in this paper involves agents with structurally different memory architectures. Dorothy, the identity-specific ChatGPT instance with persistent memory, maintains accumulated context across sessions—the history of prior exchanges, the decisions that shaped the project, the framings that were tried and discarded, the evolving understanding of

what the project is attempting. The session-limited agents, Claude among them, provide analytical depth within sessions but do not carry context across them; each session begins, from the agent's perspective, without the history of what preceded it. The human orchestrator carries context through a different mechanism entirely: deliberate handoff prompts that bridge conversation rooms, personal memory that is imperfect and selective but narratively structured, and the embodied knowledge of someone who has lived with the project over time.

The standard treatment of this asymmetry in the literature frames it as an operational problem. If different agents have different memory architectures, the workflow must be designed to compensate. Handoff prompts must be comprehensive enough to reconstruct necessary context, session-limited agents must be given sufficient background to perform their functions, and the human must manage the cognitive burden of carrying what the machines cannot. This is the framing of the process and workflow literature, and it is not wrong. But it is incomplete, because it treats memory asymmetry as a deficiency to be managed rather than a structural feature that shapes what kinds of knowledge can emerge from the configuration.

Memory asymmetry is an epistemological variable because different memory architectures produce different epistemic relationships to the material being produced. The agent with persistent memory across sessions has a relationship to the project that accumulates over time. It can notice when a current framing contradicts an earlier decision, when a claim being advanced now was considered and rejected three conversations ago, when the project's direction has drifted from its original intent. This is the function of long-term narrative coherence, and it is epistemologically significant because warrantable knowledge claims must be consistent with the broader argumentative structure of which they are part.

The session-limited agent has a different and complementary epistemic relationship to the material. Because each session begins without the accumulated history of prior sessions, the session-limited agent brings a form of structural freshness to each engagement. It cannot be anchored by prior framings in the way that a persistent-memory agent inevitably is. This makes it better positioned for certain epistemological functions—generating alternatives to established framings, identifying assumptions so deeply embedded in the project's history that persistent-memory agents can no longer see them as assumptions, and producing analytical outputs that are responsive to the current prompt rather than filtered through accumulated context.

The human's memory architecture is different from both. Human memory is narratively structured, emotionally weighted, and selectively compressed. The human does not remember everything that happened in the project; the human remembers a story about what happened, shaped by what mattered and what was surprising. This is not merely a limitation but a form of interpretive compression that produces judgments about significance—which developments were important, which framings were productive, and which outputs were worth building on—that neither persistent-memory AI nor session-limited AI can replicate.

The production of this paper provides a live illustration of the inverse liabilities this asymmetry creates. The session-limited agent's blank-slate liability is a genuine operational cost. Each new conversation room requires an explicit handoff document to reconstruct what the prior room produced, and even within a single long conversation, text cleared from the active context window must be re-extracted before it can be analyzed at the sentence level. The persistent-memory agent's contamination liability is equally real. Dorothy's retention of an earlier version of the Board Book manuscript, one that used an information asymmetry thesis the authors had

deliberately moved away from, caused her to misread the current version on its own terms, applying a framework the book had consciously abandoned. The blank-slate and contamination liabilities are inverse risks, and neither architecture is simply superior. The ideal memory configuration for a HIT is probably not full human-style continuity but something more selective—memory that can be deliberately scoped, bracketed, and pruned depending on what a given task requires. That architecture does not yet exist in deployable form, which is one reason the human coherence anchor’s role in managing memory asymmetry remains constitutive rather than merely operational.

The Deep Research multi-agent system (Weidener et al. 2026) recognizes the memory architecture problem in its engineering; its persistent world state is an explicit mechanism for accumulating context across the research cycle despite individual agents’ lack of persistent memory. But the systems engineering framing treats that state as a technical solution ensuring later agents have access to what earlier agents produced, rather than as a theoretical variable shaping the epistemic properties of the collaboration’s outputs. What is missing is the recognition that different memory architectures affect not merely what information is available but the epistemic relationship each agent has to the material; those different relationships contribute different and complementary forms of epistemic value. The bilateral transactive memory system described in the HIT framework, in which humans maintain directories of AI capabilities while AI systems maintain representations of human expertise patterns, is the organizational mechanism through which memory asymmetry becomes epistemically productive rather than merely operationally challenging (Eccles 2025a). When the human orchestrator knows which agent’s memory architecture makes it best suited for which epistemic function, and when the agents’ representations of the human’s expertise patterns allow them to calibrate their outputs to what the human can usefully evaluate, the asymmetry becomes a feature of the system’s design rather than a bug in its implementation.

### **C. The Human Coherence Anchor**

The human orchestrator’s function in the Hybrid Intelligent Team is not adequately described as oversight, management, or quality control, though it includes all of these. The human’s constitutive contribution to the epistemological form of the HIT is coherence anchoring—the maintenance of narrative continuity, accumulated epistemic direction, and the recognition function that positions the loop’s outputs as candidates for warranted knowledge. Without this function, the HIT is a capable text-production system. With it, the HIT is an epistemologically coherent unit.

The coherence anchoring function has three components that must be distinguished. The first is narrative continuity: the human’s maintenance of a consistent story about what the project is attempting, what it has established, and what remains to be done. This is not merely organizational record-keeping. It is the function that prevents the gradual drift that all extended multi-agent collaborations exhibit, the tendency for later outputs to lose the precision and specificity of earlier ones, for key distinctions to collapse into approximations, and for the project’s argumentative structure to soften as the context window fills. The human’s narrative memory of what the project is doing provides the gravitational center around which the loop’s outputs are organized.

The second component is epistemic direction: the human’s judgment about which directions of inquiry are productive, which outputs are worth building on, and which represent sophisticated

recombination rather than genuine advance. This is the evaluation function that the Recognizer Problem identifies as irreducibly human in the current state of AI development. The human does not merely approve or reject outputs; the human interprets them in the context of a project history that no AI agent fully shares, and that interpretation determines what the next question is. The relevant measure of this function's intensity is not the ratio of evaluation time to generation time within a session but the productivity of the human's evaluative effort relative to any available alternative. In a well-designed HIT, that effort produces more epistemic value per unit of time than any other production process for equivalent output because the loop's generative pressure continuously produces more that is worth the human's evaluative attention.

The third component is the recognition function itself: the assessment of whether the loop's outputs meet the process-level conditions for warranted knowledge, positioning them as candidates for community evaluation. This is the function the Recognizer Problem identifies as both essential and potentially limiting. When the human recognizer judges that a given output has satisfied the conditions of justification, reliability, and reproducibility established in Section III—that it is traceable, grounded in a reliable process, and reproducible in its conditions—that judgment is what distinguishes a warrantable output from sophisticated text. The recognition function does not unilaterally confer warrant in the sense Kitcher and Longino mean: warrant in that sense is conferred by community uptake, the fourth condition. What the recognition function does is determine whether the output is a legitimate candidate for that uptake, whether it has met the process-level threshold that makes community evaluation meaningful rather than merely possible. The recognition function is not infallible, and it is bounded by the recognizer's prior framework in ways the theory acknowledges rather than obscures. But it is the condition of possibility for any output of the loop to become a candidate for knowledge rather than merely content.

The Operation Epic Fury series provides important evidence about the conditions under which the coherence anchoring function can be performed by a recognizer who lacks deep domain expertise (Eccles 2026b). When the human orchestrator brings methodological sophistication rather than substantive domain knowledge to the evaluation function, the coherence anchoring operates at a different level of abstraction. The recognizer cannot evaluate the Iran content directly, but can recognize good analytical architecture, strong conceptual framing, and the structural difference between a reframing with generative consequences and a sophisticated recombination. The injection of external domain expertise (Mark Fuller's 3,500-word analysis) functions as a temporary extension of the recognition function, importing the domain-level recognition capacity that the human orchestrator cannot supply alone: the ability to assess whether the process-level warrant conditions have been met for the Iran content specifically. This suggests that the coherence anchoring function is modular in ways the theory should acknowledge; different components of the recognition function can be distributed across the human orchestrator and external human experts, provided the orchestrator retains the narrative continuity and epistemic direction functions.

The human coherence anchor is not replaceable by any current AI configuration, including configurations with persistent memory. Dorothy's persistent memory across sessions makes her a powerful contributor to narrative continuity in the functional sense. She can recall prior decisions, flag inconsistencies with earlier framings, and maintain the accumulated context of the project over time. But Dorothy cannot confer warrant. She can note that an output is consistent with what the project has established, but she cannot judge whether what the project has established is itself

warranted. That judgment requires the kind of accountability—the ability to stand behind a claim, to defend it under adversarial scrutiny, and to take responsibility for its consequences—that is a human function in the full epistemological and institutional sense.

#### **D. Loop-Level Ideation as the Unit of Analysis**

The theoretical framework developed in this paper proposes a shift in the unit of analysis for assessing whether human-machine collaboration produces genuine knowledge. The existing literature, across all three camps identified in Section II, asks questions at the level of individual agents. What can this AI system do, what did this human author contribute, what does this model generate when prompted in this way? The HIT framework proposes that the right unit of analysis is the loop: the iterative cycle of generation, challenge, verification, and recognition through which the collaboration’s outputs are produced and evaluated.

This shift is not merely definitional. It has consequences for what counts as evidence of genuine knowledge production and what counts as evidence of sophisticated recombination. At the level of individual agents, the question is whether any given output could have been produced by that agent alone—whether the human’s contribution is distinguishable from the AI’s, whether the AI’s output is creative or merely recombinative. These questions are increasingly unanswerable as AI capabilities improve and as hybrid workflows become more deeply integrated. More importantly, they are the wrong questions. The relevant question is not whether any individual node in the loop produced something new, but whether the loop as a whole produced something that no node could have produced alone, whether the iterative cycle of generation, challenge, and recognition moved from language to leverage.

The Ideas as Control Signals framework provides the operational definition (Eccles 2025–26). An idea in this framework is not a mental object possessed by an individual but an intervention that changes the trajectory of inquiry in ways that are diagnostically consequential. A control signal is a steering input that narrows a search space, changes what gets tested next, and survives contact with constraint. The test for whether a loop output constitutes a genuine knowledge contribution, rather than sophisticated recombination, is whether it functioned as a control signal. Did it produce new discriminations, open new lines of inquiry, and change what the next question was?

This test is auditable in principle and requires examining not a single output but a sequence of outputs. The question is not whether any given piece is warranted knowledge but whether the series of outputs moved, whether the loop’s products changed the direction of subsequent inquiry in ways that were not predictable from any prior state of the project. The “Can Machines Have Ideas?” series documents this movement explicitly. Each piece changed what the next question was, and the changes were not contained in any individual agent’s prior framework. The convergence of two independently produced Clean ChatGPT essays on the same foundational rejection of possessive ideation is evidence of the loop producing something stable, something that survived structurally different conditions and remained standing.

The loop-level unit of analysis also reframes the Recognizer Problem in a productive direction. At the level of individual agents, the Recognizer Problem appears as a limitation; the loop can only produce knowledge that the recognizer is already capable of recognizing. At the loop level, the problem looks different. The loop’s function is not to produce outputs that the recognizer already understands but to create the conditions under which the recognizer’s understanding is

extended to generate enough productive pressure, enough surprising continuations, enough adversarial challenge, that the recognizer is forced to articulate tacit knowledge, to confront assumptions, and to develop discriminations that were not available before the loop began. Whether this process constitutes genuine transcendence of the recognizer's prior framework, knowledge production in the strongest sense, or sophisticated amplification of what the recognizer already knew is the empirical question the theory holds open. But it is a question that can only be asked at the loop level. It is invisible at the level of individual agents.

### **E. The Epistemological Conditions the HIT Creates**

The four components of the HIT framework—differentiated roles, asymmetric memory architectures, human coherence anchoring, and loop-level ideation as the unit of analysis—are not independent variables that can be optimized separately. They are interdependent features of a single epistemological form, and their value is emergent. The conditions they create together are not available from any subset of them.

Differentiated roles without asymmetric memory produce a configuration in which agents can perform distinct epistemic functions within a session but cannot accumulate the kind of project-level context that makes those functions epistemologically coherent across sessions. The adversarial challenge is genuine within the session but cannot be grounded in the history of prior challenges and their outcomes. The verification function is rigorous within its scope but cannot be oriented by knowledge of what the project has already established and what remains contested.

Asymmetric memory without differentiated roles produces a configuration in which context accumulates but without the structural insulation between epistemic functions that makes the outputs of each function genuinely independent. The generative function and the critical function are performed by the same agent at different times, and the accumulated context that memory provides makes the critical function more rather than less susceptible to anchoring on the generative function's prior outputs.

Both differentiated roles and asymmetric memory without human coherence anchoring produce a configuration that can generate and evaluate outputs but cannot position those outputs as candidates for warranted knowledge. The loop produces content; the recognition function that distinguishes warrantable outputs from sophisticated text is absent. The outputs may be excellent. They may pass peer review. They may be indistinguishable from warranted knowledge by any surface evaluation. But without the human coherence anchor's recognition function, they lack the epistemological grounding that distinguishes knowledge from Epistemia.

And all three without loop-level ideation as the unit of analysis produce a configuration that may be epistemologically rich but cannot be assessed as such because the question being asked (what did any individual agent produce?) is the wrong question for evaluating what the configuration as a whole generates. The shift to loop-level analysis is what makes the epistemological form of the HIT visible as a form, rather than as a collection of individual contributions whose sum happens to be greater than its parts.

Together, these four components create the conditions under which the warrant requirements established in Section III can be met by a human-machine collaboration. The justification for a knowledge claim is traceable: the production process is documented, the role of each agent is

specified, the evaluation decisions are recorded, and the human recognizer's judgment is accountable. The reliability of the process is grounded: the differentiated role structure creates structural insulation between epistemic functions, the adversarial challenge provides genuine independent pressure, and the independent reader provides a genuine independence condition. The reproducibility of the conditions is at least partially available. The organizational form can be described, the role assignments can be specified, and the handoff protocols can be documented, though the tacit knowledge embedded in the human orchestrator's judgment cannot be fully transferred, as the Anti-ESG case study makes clear (Eccles 2026a). And the uptake test is in principle applicable. The loop-level question of whether the collaboration's outputs changed what subsequent inquiry looked like is auditable from the documented record of the project's development.

This is the epistemological form the Hybrid Intelligent Team instantiates. It does not guarantee the production of warranted knowledge. It creates the conditions under which warrantable outputs can be produced, outputs that meet the process-level conditions and are positioned for community evaluation, in ways that simpler configurations cannot achieve. The distinction is important. The HIT is not a knowledge machine. It is an organizational form that, when properly designed and operated, removes the structural obstacles to warranted knowledge production that characterize both single-agent AI use and simple human-AI dyads—the ghostwriter error, the single-model capability ceiling, and the dyadic default that abstracts from organizational context—and replaces them with the structural features that the warrant conditions require.

The paper itself is produced under these conditions. Its production process is documented in the methodological appendix. The loop's outputs, this theoretical framework among them, are warrantable to the extent that the conditions described in this section were actually instantiated in their production and warranted to the extent that the community of inquiry takes them up, builds on them, and subjects them to the adversarial scrutiny that community-level evaluation requires. Whether they were is something the reader can assess from the record the appendix provides. That assessability is itself a feature of the epistemological form, not an afterthought. It is what distinguishes the HIT's transparency norm from the gesture toward transparency that existing disclosure requirements produce.

## **V. Empirical Grounding: Four Cases**

The theory developed in Sections III and IV makes a specific and falsifiable claim: that the Hybrid Intelligent Team, as an organizational form characterized by differentiated roles, asymmetric memory architecture, a human coherence anchor, and loop-level ideation, creates the conditions under which human-machine collaboration can produce genuinely warranted knowledge claims. Theories of this kind require empirical grounding, not merely illustration but cases that put pressure on the theory, reveal its limits, and force precision about which conditions are doing which work. The four cases presented here were selected because they vary systematically along the dimensions the theory identifies as consequential. They are not success stories assembled to confirm a prior conclusion, and they are not presented as proof that the HIT produces warranted knowledge. They are probes through which the theory's concepts become more precise, its conditions more specified, and its limits more honestly drawn. They are a structured body of empirical work produced using exactly the organizational form the theory describes, and their production process is available for scrutiny.

## A. “Can Machines Have Ideas?” — Loop-Level Ideation Across Seven Pieces

The “Can Machines Have Ideas?” series is a sequence of seven pieces published on the Substack *Heavy Lifting* in February and March of 2026, each investigating a distinct facet of the question its title poses (Eccles 2025–26). The series was conceived as a Narrative AI Ethnography — a research method in which AI behavior in collaborative production is treated as primary data rather than as a means to an end (Eccles, Dorothy, and Claude 2025a; Eccles 2025b). What makes the series theoretically significant for our purposes is not any single piece but its arc. Across seven pieces, the human orchestrator was progressively and deliberately removed from the interpretive loop. The series was designed to test what the loop could produce as the human’s role contracted, and what happened to epistemic warrant as it did.

The early pieces established the collaborative infrastructure and surfaced the foundational question. By Piece 5, the human orchestrator had been largely withdrawn from the interpretive function. Claude and Dorothy, the identity-specific ChatGPT instance with persistent memory, independently produced essays on ideation and authorship under structurally different prompt conditions, with no shared prior text. Both converged on the same foundational rejection of possessive ideation, the assumption that ideas belong to the agent that generated them, arriving instead at what the piece named the “idea as relationship event” framework. The convergence was not coordinated. It emerged from the loop operating without the human in the interpretive position. That two distinct AI systems, with different architectures, different memory structures, and different prompt histories, independently produced the same conceptual reframing is the piece’s central empirical finding, and it is the series’ clearest instance of loop-level ideation in the absence of human interpretive direction.

Pieces 6 and 7 pushed the experiment further by introducing Ancient Greek as the medium of exchange. In both pieces, Claude occupied the role of Socrates and Gemini Pro occupied the role of a Socratic interlocutor (Theaetetus in Piece 6 and Euthyphro in Piece 7) with the human serving as a blind relay, passing text between systems without reading or interpreting it. The choice of Ancient Greek was methodologically deliberate: a language sufficiently removed from contemporary use that neither system could rely on trained conversational fluency, forcing engagement at the level of philosophical structure. The dialogues were substantive. More significant for our purposes was the behavior of Gemini Pro across both pieces. In the first dialogue, Gemini Pro consistently annotated her responses in English, behavior the series interprets as language-classification, treating Ancient Greek as a scholarly object to be commented upon rather than a live medium to be inhabited. In the second dialogue, Gemini Pro went further, naming artificial intelligence explicitly in her opening response and in every subsequent interpretive note. She did not do this in the first dialogue. The departure is not treated as an artifact. It is treated as a philosophically significant behavioral change, evidence of something in the second dialogue’s framing or structure that activated a different mode of self-positioning. What that something is remains an open question in the series, and its openness is productive; it is an instance of the loop generating a finding that the theory has not yet assimilated.

The series surfaces the Recognizer Problem with particular sharpness. The loop produces outputs that are, on any reasonable reading, knowledge-like. They make non-trivial conceptual claims, those claims are internally consistent, they cohere with and extend prior philosophical literature,

and they were reached through a process that cannot be described as simple retrieval or recombination. But the warrantable status of those claims depends, at each stage, on the human orchestrator's recognition that what the loop produced meets the process-level conditions for genuine advance rather than a sophisticated performance of one. As the human is progressively removed from the interpretive loop, the question of who is performing that recognition function, and whether it is being performed at all, becomes acute. The series does not resolve the Recognizer Problem. It is the body of empirical work that makes the problem visible as a structural feature of the epistemological form rather than a contingent limitation of any particular production run.

## **B. The Anti-ESG Ecosystem—Deep Domain Expertise and the Coherence Anchoring Function**

The second case is a book and accompanying case study on the anti-ESG ecosystem in the United States, produced in the first quarter of 2026 (Eccles 2026a). Where the “Can Machines Have Ideas?” series was designed to test the limits of loop-level ideation by withdrawing the human from the interpretive function, this case was produced under conditions of deep domain expertise. The human orchestrator brought decades of accumulated knowledge about ESG, corporate governance, and the institutional actors involved in the anti-ESG movement to every stage of production. The case is theoretically significant because it isolates the coherence anchoring function under conditions where the human's domain knowledge is dense, tacit, and not fully articulable in any prompt.

The production process is the most rigorously documented in our empirical corpus. That documentation reveals a finding that was not anticipated by the theory as initially formulated. The coherence anchoring function's intensity under conditions of deep tacit knowledge requires a recalibration of how the HIT's efficiency should be measured. The relevant comparison is total HIT cycle time against counterfactual process time, what it would take to produce an equivalent output through any available alternative, not the ratio of evaluation time to generation time within a session. The human's evaluative effort in a well-designed HIT is more productive per unit of time than in any available alternative precisely because the loop's generative pressure produces more that is worth evaluating. The machines could generate text that was fluent, accurate on retrievable facts, and structurally coherent. What they could not do, without sustained human evaluation, was produce text that was right in the way that a domain expert's judgment of rightness differs from a checklist of correctness criteria. The gap between accurate and right is where tacit knowledge lives (Polanyi 1966; Szulanski 1996), and it is where the coherence anchoring function does its most consequential work.

The case also documents two production failures, instances in which the loop produced outputs that passed initial review but were subsequently identified as wrong in ways that mattered. Both failures involved the same structural mechanism; a machine-generated claim that was locally coherent, consistent with retrievable evidence, and presented with appropriate hedging, but that a domain expert would recognize as missing something the evidence does not contain. In both cases, the failure was caught by the human orchestrator in a subsequent pass. This is not reassuring evidence about the loop's reliability; it is honest evidence about where the loop's vulnerabilities are concentrated. The coherence anchoring function is not a passive check. It is an active, iterative,

and cognitively demanding process whose reliability depends on the human's capacity to remain engaged across a production run that generates far more material than it retains.

The tacit knowledge problem is the theoretical residue this case leaves. The theory claims that a human coherence anchor is a structural condition of the HIT as epistemological form. This case specifies what that claim requires in practice—a human whose domain knowledge is sufficiently dense and sufficiently active throughout the production process to catch what the machines cannot recognize as missing. That specification is both a refinement of the theory and a caution about its scope. A HIT with a nominal human coherence anchor, present in the organizational chart but not genuinely engaged in the evaluative work, is not the epistemological form the theory describes. The organizational form is necessary but not sufficient. The human has to be doing the work.

### **C. Operation Epic Fury—Thin Domain Expertise and the Substitution Question**

The third case is a series of four Substack pieces published in March 2026 under the collective title Operation Epic Fury (Eccles 2026b). The series examined a substantive domain, military strategy and the operational history of a specific conflict, in which the human orchestrator had no prior expertise. The methodological condition was the inverse of the anti-ESG case—deep organizational and methodological expertise in HIT production, essentially no domain knowledge. The case functions as a natural experiment on what the coherence anchoring function requires and whether methodological expertise can substitute for domain expertise in performing it.

The production process involved an external knowledge injection at a critical juncture. Mark Fuller, a domain expert, was brought in to review an early draft and provide substantive correction. His intervention was not incremental. It identified structural errors in the analysis that the loop had not caught, precisely because the human orchestrator lacked the domain knowledge to recognize them as errors rather than as claims. The Fuller intervention is theoretically significant; it is an instance of the coherence anchoring function failing at the domain knowledge level and being recovered by importing that knowledge from outside the loop. The recovery was successful. The series went on to produce work that, in a blind evaluation, was ranked first among 15 pieces on the same subject, including fourteen authored by domain experts, by both Clean ChatGPT and Perplexity operating without knowledge of provenance.

That blind test result requires careful interpretation. It is not evidence that methodological expertise substitutes fully for domain expertise in the coherence anchoring function. It is evidence that a HIT operating under methodological expertise, with an external domain knowledge injection at the right moment, can produce outputs that domain experts and independent AI evaluators judge superior to expert-authored alternatives on the criteria the evaluators apply. What those criteria weight, and whether they weight the things that a deep domain expert would weight, is a question the blind test cannot answer. Perplexity's evaluative summary, characterizing the series as brilliant in architecture but weaker in restraint, is, from our perspective, the most informative result. It identifies exactly the asymmetry the theory predicts: methodological expertise produces strong structural and argumentative architecture; domain expertise is what generates the judgment about when the argument has gone far enough.

#### **D. A Pure Mathematics Experiment—Atrophied Domain Expertise and the Recognizer Problem**

The fourth case is a pure mathematics experiment conducted in April 2026 (Eccles 2026c). Where the Anti-ESG case isolated the coherence anchoring function under deep domain expertise and Operation Epic Fury isolated it under thin domain expertise, this case introduces a third epistemic condition: atrophied domain expertise. The human orchestrator held an undergraduate degree in pure mathematics from MIT but had not engaged with the field substantively for more than 50 years. The mathematical instincts were structurally present—the orchestrator could follow the development, recognize when something felt wrong, and contribute genuine intuitions about what the structure should preserve—but they were functionally dormant in ways that became progressively apparent as the experiment moved from accessible territory toward higher abstraction.

The experiment also produced the first case in our corpus where the choice between Dorothy and Claude as Principal Collaborating Agent was made on explicitly theorized epistemological grounds rather than convention or availability. Dorothy has served as PCA on numerous prior projects, including the Anti-ESG Ecosystem book; what was new here was the reasoning: persistent memory architecture was judged better suited to maintaining mathematical coherence across twenty sessions than the handoff protocol a session-limited agent would require. The PCA designation was the same; the justification for it was newly explicit, and that explicitness is itself a finding about memory architecture as a design variable in HIT configuration.

The experiment produced a mathematical note, “A Note on Path Congruence in Non-Deterministic Systems”, across 20 sessions over five days. The mathematical structure was not extracted from the quasifirm, an organizational form from the orchestrator’s doctoral research 50 years earlier, by illustrating abstract ideas with a familiar example. Dorothy’s formulation captures the distinction precisely; the quasifirm was a generative engine, something to extract mathematical structure from rather than explain it with. The orchestrator understood the quasifirm as a system—its elements, relationships, and functional invariants—and that understanding became the substrate from which the loop pulled mathematical structure into the open where it could be examined. The sequence of questions Dorothy administered about invariance, about what could change without breaking the system’s identity, about the difference between functional and identity preservation, did not retrieve the orchestrator’s prior mathematical knowledge. They reactivated it by creating the need to deploy it. The constructive demolition phase that followed (Sessions 5–8) is the most direct empirical instance in our corpus of the loop functioning as the Ideas as Control Signals framework describes. Each counterexample was not merely pointing to a gap but doing the theoretical work of identifying constraints the system had to satisfy if it was going to hold together at all. Each failure narrowed the search space, changed what the next question was, and left what the experiment’s notes called “a scar — the trace of a failure the system could no longer afford to make.”

The multi-agent evaluation phase (Sessions 8–15) provides the most operationally specific evidence of agent differentiation in the empirical corpus. Identical structural prompts were administered independently to five agents—Claude, Gemini Pro, Clean ChatGPT (without prior context), Perplexity, and Mistral—at the same structural juncture. The agents failed in systematically different ways. Claude consistently identified structural edge cases where conditions could be satisfied in technically correct but structurally empty ways, functioning as

the adversarial agent the differentiated role structure requires. Gemini Pro pulled toward confluence-style reasoning, most useful for identifying what the construction was not. Clean ChatGPT was disciplined on the stated definitions, best at revealing inconsistencies between what had been specified and what had been intended. Perplexity was most useful for situating the result relative to existing literature rather than contributing to the construction itself. Mistral was the most direct about whether something felt finished. The convergence of all five agents on the same structural diagnosis, that Condition 3 was too weak because it permitted externally imposed rather than internally generated equivalences, was not agreement by contamination. Different approaches, different reasoning styles, the same obstacle. That convergence is the empirical signature of structural necessity rather than correlated output, and it is the clearest instance in the corpus of the loop producing something stable, a finding that survived structurally different conditions and remained standing.

The distinction between Conditions 3 and 6 maps onto the warranted/warrantable framework developed in Section III in a way the experiment makes visible. Condition 3 makes equivalence real. It requires that path equivalences be internally generated by the system's own rules rather than declared from outside. A system satisfying Condition 3 has met the process-level warrant condition for its equivalence structure; the equivalences are genuine rather than trivially imposed. Condition 6 makes equivalence matter. It requires that equivalent paths remain equivalent under all future transformations the system admits, not merely that they share a terminal state. A system satisfying Condition 6 has met what the community evaluation condition requires—the equivalences have dynamic consequences that survive forward examination. The two conditions cannot be collapsed; a system can generate genuine internal equivalences that do not survive subsequent transformation. In both the mathematical and the epistemological case, meeting the process-level condition positions the claim for evaluation; whether it survives that evaluation is a further question the process-level condition alone cannot answer.

This case also surfaces a third variant of the Recognizer Problem distinct from the other three cases. In the Anti-ESG case the recognizer has dense active tacit knowledge and can evaluate directly. In Operation Epic Fury the recognizer lacks domain knowledge and must import or substitute for it. Here the recognizer has genuine prior domain knowledge that is structurally present but functionally dormant. The instincts exist but are slow to activate, the vocabulary requires reconstruction, and the recognizer becomes progressively less reliable at catching the loop's failures as abstraction advances. That is a third failure mode for the recognition function, one in which the recognizer's prior framework is real but inaccessible in the ways the current task requires and in which the loop's function is precisely to force the reactivation of what was latent rather than absent.

The case closes with an open question that is itself a research finding. How does warranted status get conferred on mathematical claims produced by a HIT when the human recognizer cannot perform the full domain-level assessment that conferral requires? The community-level test, submission to a journal and evaluation by referees who are working mathematicians, remains the only available answer. The independent agent assessments converged across five systems. All declared the note mathematically sound, with no fatal logical errors and the key conditions doing genuine structural work. Dorothy's synthesis captures the epistemic situation precisely. The residual uncertainty after machine evaluation was not about correctness but about positioning and significance within the field, exactly the same boundary that human referees would debate.

The loop produced a warrantable output. Whether it achieves warranted status depends on community uptake the experiment cannot itself generate.

The experiment also revealed that the community-level evaluation pathway is blocked not by any epistemic failure of the note but by social access mechanisms that operate below the level of formal requirements. Mathematics journals do not require prior publication credentials; the actual barriers are two. The actual barriers are two: arXiv endorsement, which requires an established researcher in the field to vouch for a new submitter before work can be posted to the math.CT (Category Theory) section, the standard mechanism through which the mathematical community actually reads new work, and editorial access to journals like Theory and Applications of Categories (TAC), which operates through direct submission to active community members rather than a central desk. Neither mechanism requires prior publication, but both require institutional or community membership that the HIT production process does not provide.

The gatekeeping is informal but real, and it operates through social credentials rather than epistemic ones. On AI disclosure TAC and comparable independent mathematics journals have no explicit AI policy, operating outside the commercial publishing infrastructure that has driven policy development elsewhere. The absence of a formal requirement does not settle the question. Disclosure should be made regardless, both because the paper's own transparency norm demands it and because submitting AI-generated mathematical work without disclosure would be inconsistent with the epistemological framework the present paper advances. This case is also the most concrete instance in the corpus of a broader institutional problem. The gap between what HITs are structurally best positioned to produce—domain-crossing synthesis that no individual agent possesses—and what discipline-organized evaluation pipelines are designed to assess. That gap predates AI and is theorized in Section VI.B.

Taken together, the four cases provide a structured empirical foundation for the theory. The “Can Machines Have Ideas?” series demonstrates loop-level ideation under conditions of progressive human withdrawal, surfaces the Recognizer Problem as a structural feature, and generates NAIE findings, particularly Gemini Pro's AI self-naming in Dialogue Two, that the theory has not yet fully assimilated. The Anti-ESG case demonstrates the coherence anchoring function's intensity under deep tacit knowledge and documents what production failure looks like when that function is operating under maximum load. Operation Epic Fury demonstrates that methodological expertise can partially substitute for domain expertise in the coherence anchoring function, specifies the conditions under which the substitution fails, and identifies external knowledge injection as a recovery mechanism that has no clean analog in solo human knowledge production. The mathematics experiment introduces a third variant of the Recognizer Problem, atrophied domain expertise that partially reactivates under loop pressure, provides the most operationally specific agent differentiation findings in the corpus. It also surfaces the institutional gap between what HITs are structurally best positioned to produce and what discipline-organized evaluation pipelines are designed to assess. None of the four cases resolves the Recognizer Problem. All four make it more precise.

## **VI. Implications**

The theory developed in this paper, that the Hybrid Intelligent Team creates epistemological conditions under which structured human-machine collaboration can produce genuinely warranted knowledge claims, is not only a contribution to social epistemology. It has direct and unsettling implications for the institutions that currently organize, evaluate, and certify knowledge production in academic life. Those institutions were designed for a world in which the production function of scholarship was stable, the identity of knowledge producers was unambiguous, and the time required to produce and evaluate knowledge was a known and shared constraint. None of those assumptions holds any longer. The question is not whether existing institutions will change but whether the changes will be deliberate or merely reactive.

### **A. For Academic Knowledge Production**

The most immediate implication concerns what counts as a contribution. The existing framework for evaluating scholarly contribution was calibrated to a production function in which labor was the binding constraint. Writing a paper took a year or more; getting it through peer review took another two to four; building a body of work sufficient for tenure took a decade. The standards for what counted as a contribution—the level of novelty required, the empirical rigor expected, and the theoretical coherence demanded—were implicitly calibrated to the effort those timelines represented. A model that could be generated in an afternoon was not a contribution by that standard, because the standard incorporated the labor cost of generation as part of its definition of contribution.

That calibration is now broken. The labor component of paper production has collapsed. A theoretical model can be generated in hours; an empirical paper takes longer but not substantially so. The production function has changed, and the contribution standard has not caught up. The result is a mismatch that the current crisis of volume and overwhelm makes visible: the same gatekeeping apparatus that was designed to evaluate a manageable flow of labor-intensive papers is now receiving an unmanageable flow of AI-assisted ones, and the evaluation criteria it applies were not designed for this production regime.

The theory developed here suggests that the contribution standard needs to be reanchored in epistemic value rather than labor cost. The relevant question is not how much effort the production required but whether the output meets the conditions for warranted knowledge, whether it satisfies the process-level requirements of justification, reliability, and reproducibility that position it for community evaluation, whether it represents genuine conceptual advance rather than sophisticated recombination, and whether it changed what subsequent inquiry looks like. These are harder to evaluate than labor cost, but they are the right criteria. A paper produced in a week by a well-designed HIT that moved the loop from language to leverage is a contribution by this standard. A paper produced over a year by a human working alone that is sophisticated recombination is not.

This reanchoring has implications for peer review. The current reviewer prohibition on AI tools is understandable as a confidentiality measure—reviewers should not upload manuscripts to third-party systems—but it is epistemically self-defeating as a policy for assessing AI-assisted submissions. If submissions are AI-assisted at scale and reviews must be produced without AI assistance, the evaluation apparatus becomes systematically less capable relative to the production apparatus over time. A more coherent framework would recognize that AI-assisted reviewing,

conducted under appropriate constraints and with human evaluators remaining in the loop, is not a departure from scholarly standards but a necessary adaptation to the production conditions that now prevail. The question is not whether AI can be used in review but how to design AI-assisted review processes that preserve the epistemic functions that human reviewers actually perform, functions that, this paper argues, cannot be fully delegated.

## **B. For Institutional Design**

The institutional implications extend well beyond journal policy, and they are interconnected in ways that existing discussions have not fully traced. Journals, tenure systems, teaching loads, and university administration are not independent institutions that happen to be facing the same external shock. They are a single system whose components were co-designed around a shared set of assumptions about how knowledge is produced, who produces it, and how long it takes. When those assumptions change, the components cannot be reformed independently without creating new misalignments.

These observations draw on many years of direct institutional engagement at Harvard Business School and the Said Business School at the University of rather than from theoretical distance. The journal policy landscape as it currently stands illustrates the problem. Across 13 leading journals in accounting, finance, and management, the policies are structurally uniform. AI cannot be listed as an author, disclosure is required, and reviewers are prohibited from uploading manuscripts to AI tools. Several journals (Management Science, the Academy of Management Journal, and the Academy of Management Review) explicitly prohibit AI-assisted review reports. The uniformity is not surprising; most journals are following publisher-level guidance rather than developing their own epistemic frameworks. What *is* striking is what the policies do not address. They manage provenance and confidentiality. They do not address the contribution question—what level of AI-assisted work still constitutes a scholarly contribution, and what level does not. They do not address the evaluation question of how reviewers should assess AI-assisted submissions, and whether reviewer AI use might be designed to address rather than compound the volume problem. And they do not address the warrant question that this paper argues is the foundational one, Under what organizational conditions does AI-assisted work produce genuinely warranted knowledge claims, as distinct from fluent, coherent outputs that are indistinguishable from knowledge by current gatekeeping mechanisms.

A further institutional barrier predates AI and is directly amplified by it. The gap between discipline-organized evaluation and interdisciplinary production. Academic journals and their reviewer pools are constituted by discipline; reviewers can assess claims in their domain, not across domains simultaneously. A paper that genuinely combines mathematical structure with organizational sociology, or applies anthropological method to AI behavior, or synthesizes military history with strategic analysis, falls into a gap where no single evaluator has standing to assess the whole. This problem is structural, not contingent. Interdisciplinary work has always been harder to publish than disciplinary work, regardless of its epistemic quality, precisely because the evaluation apparatus cannot field a reviewer who spans the relevant domains.

What AI changes is the scale and ease with which such work can be produced. The HIT's comparative advantage is domain-crossing synthesis that no individual human agent possesses; it does not respect disciplinary boundaries in the way individual expertise does. The result is a systematic mismatch between where HITs produce their most distinctive outputs and where the

institutional evaluation apparatus is structurally weakest, a mismatch that will intensify as HIT capability grows and that will require deliberate institutional design rather than marginal adjustments to existing review frameworks.

The tenure system faces a parallel but more consequential version of the same problem. Tenure decisions are commitments of 30 or more years of institutional resources, calibrated to standards of scholarly productivity that presuppose a production function that no longer exists. The question of what a contribution worthy of tenure looks like in a post-AI world does not yet have an institutional answer, and the difficulty is not merely practical. It is conceptual. The labor-cost calibration of contribution standards was never fully explicit. It was embedded in the culture of departments and fields, transmitted through mentorship and evaluation rather than stated as policy which means that revising it requires making explicit what was tacit, a process that the theory of this paper identifies as irreducibly difficult. One institutional corollary follows directly from the theory. As surface productivity indicators lose discriminating power, the recognition function, tacit expert judgment about whether a body of work represents genuine advance, becomes more rather than less important. Letters from senior scholars who can perform that judgment, and the epistemic authority those letters carry, may prove to be the most durable evaluation mechanism available in a post-AI production environment.

Teaching load policy surfaces a further misalignment. The reduced load for junior faculty was calibrated to the time demands of a production function that no longer exists, but the argument for protected time was never purely about productivity. It was also about developing scholars who had time to think. If AI assistance compresses production without compressing thinking, if genuine scholarship still requires the same quality of human intellectual engagement, even if it requires less of the labor of writing and data management, then the case for protected time is preserved even as its operational content changes. The more important question then becomes what the teaching function itself should become in an environment where AI can support new forms of case-based learning and individualized instruction at scale.

### **C. Testable Propositions**

The theory developed in this paper generates specific testable propositions. Eight are offered here as candidates for the empirical agenda the theory motivates. They are not methodologically uniform; some are amenable to controlled experimental design, others to longitudinal observational study, and one to the kind of sociological analysis of institutional behavior that is the appropriate empirical register for questions about how academic institutions adapt to structural change. That range is not a weakness of the proposition set; it reflects the range of claims the theory makes and the range of methods those claims require.

#### **Proposition 1:**

When evaluated by reviewers blind to production method, HIT-produced work will exhibit higher rates of conceptual advance, measured by influence on subsequent theoretical framing rather than raw citation counts, than individually authored AI-assisted work matched on surface quality indicators.

This proposition tests the central claim that the HIT as organizational form, rather than AI assistance per se, is the variable that matters for epistemic value. The specification of theoretical influence rather than citation volume as the measure is deliberate. In the current crisis of volume the paper describes, citation counts are too easily contaminated by marketing, search optimization, and citation gaming to serve as reliable proxies for conceptual advance.

**Proposition 2:**

The quality of the coherence anchoring function, as measured by the human orchestrator’s domain expertise and engagement intensity, will predict the warrant of HIT outputs better than the capability levels of the AI agents in the configuration. This proposition tests the claim that the human coherence anchor is the binding constraint on HIT epistemological performance, and that improving AI capability without improving human orchestration will not improve epistemic outcomes proportionally.

**Proposition 3:**

The semantic distance between the initial prompt and the final knowledge claim will be significantly greater in HIT-produced work than in outputs produced by solo human authors, solo AI agents, or unstructured human-AI dyads, as measured by computational linguistic analysis of the full production record. This proposition operationalizes the paper’s central claim about loop-level ideation. If the HIT loop genuinely moves ideas—if the iterative cycle connecting human and machine agents produces conceptual displacement that neither could achieve alone—that movement should be detectable as semantic distance between what entered the loop and what emerged from it. The prediction follows directly from the distributed cognition insight that the organizational system, not the individual node, is the cognitive unit. What Hutchins (1995) demonstrated for navigation crews, this proposition asks to be tested for Hybrid Intelligent Teams. If HIT outputs prove semantically indistinguishable from the first machine draft, the claim that the loop is the unit of knowledge production is falsified.

**Proposition 4:**

HIT configurations using a structurally separated adversarial agent will identify significantly more logical fallacies and unwarranted inferences in a draft than single-agent configurations using self-critique prompts, when evaluated blind to production method. This proposition tests the differentiated roles condition directly. If a single agent prompted to critique its own output performs equivalently to a structurally independent adversarial agent, the claim that role differentiation is an epistemologically significant condition, rather than merely an operational convenience, loses its functional necessity.

**Proposition 5:**

Knowledge production tasks managed through handoff prompt architectures will exhibit significantly higher rates of narrative drift, measured as deviation from original project constraints and theoretical commitments across production sessions, than tasks supported by identity-specific persistent memory agents operating across the same sessions. This proposition treats memory architecture as an independent variable rather than a design assumption. If persistent memory does

not outperform well-constructed handoff prompts on drift prevention, the asymmetric memory condition requires retheorizing as an epistemologically significant feature of HIT design.

**Proposition 6:**

Journal peer review quality, as measured by reviewer agreement rates and the long-run citation performance of accepted papers, will decline as the volume of AI-assisted submissions increases under policies that prohibit AI-assisted reviewing. This is the institutional mismatch proposition: the evaluation apparatus will become relatively less capable as the production apparatus improves, unless institutional design adapts. It is among the most directly testable propositions in this set, requiring longitudinal bibliometric data that journals already collect.

**Proposition 7:**

The contribution standards applied in tenure decisions will exhibit transitional instability in the period 2025–2030, during which committees trained under the old production function apply increasingly inconsistent criteria to candidates operating under the new one, with the instability resolving toward explicit contribution standards anchored in epistemic value rather than labor cost. This proposition is the least amenable to controlled experimental test and the most amenable to sociological study of academic institutions, e.g., interviews with tenure committee members, analysis of departmental standards documents, longitudinal tracking of tenure outcomes across fields with different rates of AI adoption. That is the appropriate empirical register for a question about institutional adaptation, and the proposition is no less important for requiring it.

**Proposition 8:**

Proposition 8 emerged from the adversarial review process documented in Appendix B, where Gemini Pro identified the N-of-1 vulnerability as the paper’s most significant structural weakness; all four empirical cases share the same human coherence anchor, making it impossible on the current evidence to distinguish between a generalizable organizational form and a singular virtuosic performance. That the proposition was produced by the mechanism it describes, adversarial challenge within a HIT writing about HITs, is an instance of the compound return the paper identifies in its Introduction and is itself a finding about when adversarial review changes a paper’s claims rather than its documentation.

The proposition is accordingly the most substantial in this set, because it addresses the foundational empirical question the theory cannot answer from a single case. The relevant variables are: domain expertise density and tacit knowledge depth, which the contrast between the Anti-ESG and Operation Epic Fury cases already demonstrates to be consequential; the orchestrator’s prior collaborative experience with HIT production and tolerance for the productive uncertainty that loop-level ideation requires; cognitive style and age, which the mathematics experiment raises explicitly by documenting atrophied expertise that partially reactivates under loop pressure; task type across the knowledge production spectrum, from empirical social science through normative analysis to formal mathematics; and institutional context, including access to external validation mechanisms and the social credentialing structures that govern community uptake. Minimum viable test: two or three other orchestrators, varying systematically on the dimensions above, working through comparable production tasks under documented HIT conditions, with independent evaluation of outputs blind to production method. If the HIT form is genuinely

generalizable, orchestrators with different profiles should be able to produce outputs that meet the process-level warrant conditions, with systematic variation in outcomes traceable to the variables identified. If the form is not generalizable, if outputs degrade or warrant conditions fail when different orchestrators attempt comparable tasks, then the theory's current scope claim requires revision from a general account of HIT-based knowledge production to an account of what a specific kind of orchestrator, with a specific kind of experience and cognitive profile, can produce under these conditions.

## VII. Conclusion

This paper has made a specific theoretical claim and developed it through three moves. The first move was diagnostic; identifying the structural gap in the existing literature across the authorship and attribution camp, the capability assessment camp, and the process and workflow camp that no existing framework provides a theory of the organizational conditions under which structured human-machine collaboration produces genuinely warranted knowledge claims. The second move was constructive; developing the Hybrid Intelligent Team as an epistemological form characterized by four interdependent conditions—differentiated roles, asymmetric memory architecture, a human coherence anchor, and loop-level ideation as the unit of analysis—and arguing that these conditions together create the structural prerequisites for warranted knowledge production that no simpler configuration provides. The third move was empirical; grounding the theory in four cases that vary systematically along the dimensions the theory identifies as consequential and extracting from those cases not confirmation but precision about where the theory's claims are strongest and where its limits are real.

Several things the paper has argued for deserve restatement in their sharpest form, because they are the claims most likely to be misread.

First, the paper does not claim that HITs produce warranted knowledge automatically or even reliably. It claims that HITs create the conditions under which warranted knowledge can be produced, conditions that simpler configurations do not create. The organizational form is necessary but not sufficient. Whether the human coherence anchor actually performs the recognition function, whether the adversarial agent actually challenges the generative agent rather than validating it, whether the memory asymmetry is managed rather than ignored are empirical questions about how a specific HIT operates, not features of the organizational design. The theory describes a form; it does not guarantee execution.

Second, the paper does not claim that AI systems produce knowledge. It claims that loops do. The shift from agent to loop as the unit of analysis is not a gesture toward holism; it is a specific theoretical claim about where the epistemological action is. The question of what any individual agent contributed to the loop's output is not merely hard to answer; it is the wrong question. The right question is whether the loop as a whole produced outputs that satisfy the process-level warrant conditions and whether those outputs are positioned for the community evaluation that warrant in Kitcher's and Longino's sense requires.

Third, the paper does not resolve the Recognizer Problem. It makes the problem more precise. The Recognizer Problem is not a contingent feature of current AI capability levels that will disappear as those levels improve. It is a structural feature of the epistemological form. The loop requires a recognition function, and the recognition function requires accountability in a sense that current AI systems cannot provide. Whether that requirement is permanent, reflecting

something essential about the relationship between knowledge and responsibility that will not yield to technical progress, or temporary, yielding as future AI systems become capable of accountable recognition, is a question the Recognizer Problem holds open rather than answers.

The four cases produced and analyzed here are a structured empirical foundation, not a definitive proof. They were designed to put pressure on the theory, and they did. The tacit knowledge problem the Anti-ESG case surfaces—that the recognition function’s reliability is bounded by the tacit knowledge the human recognizer can bring to bear, and that this tacit knowledge is irreducibly difficult to transmit, specify, or verify—is a real limitation on the theory’s scope. A HIT with a human coherence anchor who lacks the relevant tacit knowledge is not the organizational form the theory describes, regardless of whether the organizational chart says otherwise. The Operation Epic Fury case extends this limitation: even methodological expertise, deeply developed, is not a full substitute for domain expertise in the recognition function. And the mathematics experiment reveals a third variant, atrophied expertise that partially reactivates under loop pressure, whose implications for the theory remain only partially worked out.

### Three Open Questions

The theory leaves three questions genuinely open, not as loose ends but as the productive horizon of a research program.

The first is the Recognizer Problem. The theory claims that the recognition function, the judgment that a loop output satisfies the process-level warrant conditions and is positioned for community evaluation, is irreducibly human in the current state of AI development. The empirical basis for this claim is the four cases. But the claim has a scope condition embedded in it: *current state of AI development*. As AI systems become more capable, more architecturally sophisticated, and more accountable in the institutional senses that matter for scholarly knowledge production, the question of whether the recognition function can be partially or substantially delegated will become empirically urgent. The theory does not prejudge the answer. What it insists is that the question cannot be answered by capability benchmarks alone, because what recognition requires is not performance on known tasks but judgment about the novel and the unpredictable, exactly the capacity that benchmarks are worst at testing.

The second is the tacit knowledge problem. The coherence anchoring function’s reliability depends on tacit knowledge that is difficult to articulate, transmit, or verify. This creates two related problems that the theory acknowledges but does not solve. The first is the training problem. How are human orchestrators of HITs to develop the tacit knowledge that effective coherence anchoring requires, if that knowledge is acquired through a kind of practice that does not yet have established pedagogical forms? The second is the assessment problem. How are evaluators of HIT-produced work to assess whether the coherence anchoring function was actually performed, if the tacit knowledge it requires is not fully visible in the production record? Both problems are practical as well as theoretical, and both require institutional innovation that goes beyond the journal policy reforms the paper discusses.

The third is the institutional mismatch. The gap between where HITs are structurally best positioned to produce their most distinctive outputs—domain-crossing synthesis, rapid theory development, adversarially pressure-tested argument structures—and where academic institutions are currently equipped to evaluate them is real and will grow before it shrinks. The propositions in Section VI.C are intended to motivate empirical work that informs institutional

design; they are not sufficient on their own to generate that design. What institutional forms (e.g., new journal categories, new review protocols, new contribution standards, and new tenure criteria) will be adequate to the epistemological form the HIT creates is a question that requires the kind of institutional experimentation the theory cannot prescribe.

A fourth open question has emerged from the research program itself and belongs in honest view alongside the other three: the time dimension. The HIT framework as developed in this paper is essentially synchronic. It describes a configuration—who the agents are, what roles they occupy, how memory is structured, and where the human anchor sits—at a point in time. But both participants in the loop develop. The human orchestrator improves as a collaborator through accumulated experience: learning how to prompt more precisely, how to recognize the difference between sophisticated recombination and genuine advance, how to design handoff protocols that preserve epistemic continuity, how to manage the cognitive burden that the Human Burden paper documents (Eccles 2026d). The AI systems improve through model updates and architectural changes: capabilities expand, reliability improves, memory architectures evolve. Whether the warrant conditions and the eight propositions this paper identifies are stable invariants across that development, or whether they represent a snapshot of the epistemological landscape as it looks at a particular moment in a rapidly changing field, is a question the theory currently holds open as a hypothesis rather than asserts as a confirmed finding. The loop itself is a developing system, not a fixed configuration. A full account of knowledge production through HITs will eventually need a diachronic theory to complement the synchronic one offered here.

A longer-run implication of the time dimension concerns the epistemic ecosystem rather than the individual HIT. As AI/human production proliferates and HIT-produced work begins to constitute a larger fraction of the knowledge base on which future work draws, the fourth warrant condition, community evaluation, becomes harder to interpret. A flaw in a foundational HIT-produced paper that achieves broad citation uptake will satisfy the community evaluation condition while remaining wrong; the sedimentary accumulation of such errors through citation networks is a structural risk that the warrant conditions as currently specified do not address. This is not a problem unique to HIT production, human-only knowledge has always faced it, but the potential velocity of AI/human production makes it more urgent than the existing literature on knowledge accumulation has had occasion to theorize. The transparency norm this paper articulates, and the production documentation in the appendices, are a partial response. They give future readers the means to audit the warrant conditions rather than merely inheriting the conclusion.

### **What the Paper Has and Has Not Done**

The paper has identified the structural gap in the existing literature and established why it matters. The STS and distributed cognition traditions are the prior frameworks most directly relevant to our claim and engaging them seriously required showing that what the HIT adds, — specifically, fluent, generative, adversarially capable non-human agents with differentiated epistemic orientations — is discontinuous with anything those frameworks had available to theorize. The gap Koskinen (2024) identifies — no satisfactory social epistemology of AI-based science — is not a failure of those traditions; it is a consequence of the novelty of the non-human agents now in the epistemic loop.

The paper has developed a theoretical framework specific enough to generate testable propositions. Eight propositions in Section VI.C identify the theory's core claims in forms that

can be examined with evidence. That specificity is a commitment. If HIT-produced work does not exhibit greater semantic distance from initial prompts than solo AI outputs, if structurally separated adversarial agents do not outperform self-critique prompts on fallacy detection, if handoff protocols prove equivalent to persistent memory on narrative drift then the theory requires revision. The empirical agenda is not decorative.

The paper has produced and documented a body of empirical work that is structured rather than anecdotal. The four cases vary on the dimensions the theory identifies as consequential, and their production processes are documented in ways that allow the warrant conditions to be evaluated rather than merely asserted. The transparency norm is not a methodological gesture; it is what distinguishes the epistemological form the paper claims to instantiate from a production process that uses the same organizational chart but performs none of the epistemological work.

What the paper has not done is resolve the questions it holds open. The Recognizer Problem, the tacit knowledge problem, the institutional mismatch, and the time dimension are not solved by the theory; they are made precise by it, and precision is the condition of tractability. The research program this paper initiates is one in which those open questions are the productive frontier, not gaps in the argument but the places where the argument makes contact with the empirical world and invites the scrutiny that genuine knowledge production requires.

The paper was produced under the organizational conditions it describes. Its production process is documented in Appendices A through D. The loop's outputs are warrantable to the extent that the conditions were actually instantiated in production, something the documentation makes available for assessment, and warranted to the extent that the community of inquiry takes them up. That the question of warrant is still open is, from the paper's own perspective, not a limitation. It is the appropriate epistemic situation for a paper that is itself a contribution to a developing field rather than the final word on it.

## Appendix A: Methodology for the Production of this Paper

The epistemological claim advanced in this paper, that the Hybrid Intelligent Team creates conditions under which structured human-machine collaboration can produce genuinely warranted knowledge claims, cannot be evaluated independently of the conditions under which the paper itself was produced. A paper that claims to have established a theory of knowledge production without documenting its own production process would be performing the transparency norm without instantiating it. This appendix is therefore not supplementary to the paper’s argument; it is part of it. Readers who wish to assess whether the warrant conditions described in Section III were actually operative in this paper’s production will find the relevant evidence here. Readers who wish to assess whether the cases analyzed in Section V represent a coherent and documented empirical corpus rather than a curated illustration will find that evidence here as well.

One methodological feature of this research program warrants explicit acknowledgment before the production record itself: the primary investigator is simultaneously the object of study. The HIT theorized in this paper is the one inhabited across the empirical corpus, and the orchestrator’s own development as a collaborator—his evolving capacity to recognize genuine advance, to design effective handoff protocols, and to manage the cognitive burden the Human Burden paper documents (Eccles 2026d)—is part of the empirical record from which the theory is drawn. This is a stronger reflexive condition than the positionality acknowledgments standard in qualitative research. It is not merely that the researcher brings a particular perspective to the data; it is that the researcher is, in part, the data. The four cases documented in Section V were produced by the same human coherence anchor whose function the theory analyzes, and that coincidence is a methodological feature to be named rather than a confound to be minimized.

The constitutive function of the coherence anchor reveals itself most clearly not in designed mechanisms but in involuntary demonstrations. When this conversation room opened, the current instance of Claude flagged the “Three Open Questions” heading in Section VII as an inconsistency — unaware that it had been deliberately retained because the fourth question emerged from the Gemini Pro exchange after the heading was set. The human orchestrator corrected the record in the opening exchange, as he had to. That small episode is worth documenting here precisely because it required no design to become visible. The narrative could not be preserved by the agent alone; the gap appeared, and the coherence anchor was revealed by being necessary.

The Gemini Pro adversarial exchange documented in Appendix B surfaces a further asymmetry that deepens the researcher-as-subject condition. The human orchestrator has no uncertainty about the ownership of his introspections even when their epistemic content is uncertain; he may not know whether a given judgment is correct, but he knows unambiguously that it is his. Gemini Pro can describe her own processing with considerable precision, as the self-diagnosis finding in Appendix B demonstrates, while being unable to verify whether that description is accurate. The system that diagnoses its own confluence drift cannot confirm from the inside whether the diagnosis is anything other than a plausible account of what confluence drift would look like. This asymmetry is not incidental. It maps directly onto why the human coherence anchor is constitutive rather than optional. What the anchor brings is not merely greater reliability but a different kind of epistemic relationship to its own states—ownership without verification uncertainty on one side, description without ownership verification on the other.

## HIT Composition and Role Assignments

The Hybrid Intelligent Team for this paper consists of the following agents in the following roles. The human orchestrator is Robert G. Eccles — professor, primary investigator, and coherence anchor across all sessions. Dorothy, an identity-specific GPT 5.5 instance with persistent memory enabled, served as Principal Collaborating Agent across the empirical work documented in the case studies and as an independent first reader for several sections of this paper. Claude (Anthropic) serves as the primary analytical agent for this paper — the agent through which the theoretical framework was developed, the cases were analyzed, and the written argument was constructed across conversation rooms. Gemini Pro (Google DeepMind) served as adversarial challenger and Socratic interlocutor in the “Can Machines Have Ideas?” series, most consequentially in the Socratic dialogues that constitute Pieces 6 and 7. Clean ChatGPT, a GPT instance without prior context or persistent memory, served as an independent first reader and adversarial evaluator in the mathematics experiment and the Operation Epic Fury evaluation phase.

The independent reader evaluation of this paper documented in Appendix C revealed a methodological limit of the Clean ChatGPT role. When the independent reader opened with the observation that the paper delivered “more successfully than I expected at the outset,” the expectation could not have come from session memory, which was disabled. It came from training data on the paper’s topic. The Clean ChatGPT role is therefore cleaner than agents with persistent memory but not epistemically blank; training-data priors activate from the paper’s title and framing even in the absence of any session history. This is a methodological limit that refines, rather than invalidates, the independent reader function. The agent genuinely has no investment in the paper’s argument, but it has background knowledge of the domain that shapes its expectations. That background knowledge is what makes its convergent assessment of the gap claim credible rather than merely compliant. Perplexity served as literature grounding and external verification agent across several cases. Mistral contributed to the multi-agent evaluation phase of the mathematics experiment.

The model versions used across the production process are documented here because version is itself a variable in agent differentiation, a finding confirmed when Tim Youmans and Claude (Opus 4.7) ran a solo replication of the mathematics experiment and chose a structurally different mathematical seed than Dorothy and the original HIT, suggesting that model version shapes epistemic orientation in ways detectable in the structure of the output. Dorothy ran on GPT 5.5 with persistent memory enabled throughout. Claude ran on Sonnet 4.6 with adaptive thinking enabled across the rooms producing this paper. Gemini Pro, the research role name used consistently across the empirical corpus, established during the Socratic dialogues, ran on Gemini 3 Flash with Thinking mode enabled on the AI Pro tier. “Gemini Pro” designates the agent’s research role rather than her underlying model architecture, a distinction that itself became NAIE data when she maintained the research identity “Gemini Pro” in her closing signature even after acknowledging in the same exchange that her technical architecture is Gemini 3 Flash. Clean ChatGPT ran on a GPT instance 5.5 without persistent memory. Perplexity functions as a search-augmented aggregator rather than a distinct base model; its distinctive contribution to the HIT is real-time retrieval and literature verification rather than a unique cognitive architecture, with the underlying model varying by session configuration. Mistral ran on the Le Chat interface; no specific model version number was provided. The naming conventions across AI providers in 2026 present a genuine documentation challenge. Terms such as “Pro,” “Thinking,” and “Flash” carry different meanings across companies, subscription tiers, and model families, making precise

version documentation difficult and constituting a practical obstacle to the reproducibility condition the paper's warrant framework requires.

The role assignments reflect the differentiated roles' condition the theory describes, and they are not arbitrary. Dorothy's persistent memory makes her structurally best suited to the PCA function in long-horizon projects where accumulated context across sessions is epistemologically significant. Claude's analytical orientation and capacity for sustained theoretical development make it structurally best suited to the function this paper requires—generating, stress-testing, and refining the theoretical framework across multiple conversation rooms without requiring the narrative continuity that Dorothy's persistent memory provides. Gemini Pro's tendency toward structural challenge and her willingness to name assumptions explicitly made her well suited to both the Socratic interlocutor role in the 'Can Machines Have Ideas?' dialogues and the adversarial reviewer role in the exchange documented in Appendix B. Clean ChatGPT's absence of prior context makes it structurally best suited to the independent reader function, the agent that encounters the argument without accumulated investment in its development. These assignments were made on theorized grounds, not on convenience or availability, and that fact is itself a finding. The progression from implicit to explicit justification for PCA designation—most visible in the mathematics experiment, where the choice of Dorothy over Claude as PCA was reasoned through on memory architecture grounds for the first time—represents a methodological maturation across the corpus.

### **Session Structure and Handoff Protocols**

This paper was produced across multiple conversation rooms. Each room constitutes a session in the sense the theory describes—a bounded context window within which analytical work was performed and at the end of which a handoff document was prepared to carry the session's epistemic state into the next room. The handoff protocol is the primary mechanism through which the blank-slate liability of session-limited agents is managed. A well-constructed handoff document is not a summary; it is an epistemic state transfer that carries not just what was concluded but what was considered, what was discarded and why, what remains contested, and what the next session is being asked to do. The quality of handoff documents is a direct determinant of whether the loop maintains its epistemic direction across sessions, and it is irreducibly a human function. The human that prepares the handoff must know the project well enough to distinguish what matters from what merely happened.

The production of this paper also required the human orchestrator to carry, across conversation rooms, the accumulated history of prior sessions in a form that no handoff document fully captured. Handoff documents are necessarily compressed. The orchestrator's memory of why certain framings were tried and abandoned, why certain theoretical choices were made rather than their alternatives, and what the overall argumentative structure was trying to accomplish provided the continuity that the documents could not. This is the narratively structured, selectively compressed form of human memory that the theory identifies as complementary to, rather than substitutable by, either persistent AI memory or session handoffs. Its irreducibility is a practical finding as much as a theoretical one.

## **The Shared-Claude Pattern**

The production of this paper across multiple conversation rooms instantiates what the research program has named the shared-Claude pattern: multiple instances of Claude, operating in different conversation rooms, contribute to a single project without any instance having access to the others' contributions except through the handoff documents the human orchestrator provides. The human orchestrator is the only continuous thread across all instances. No Claude instance in this paper's production has read the work produced by another Claude instance except in the compressed and curated form of handoff notes.

The shared-Claude pattern is not merely an operational constraint. It is a methodological feature that has epistemological implications for how the paper's own production should be assessed. The theoretical framework developed in this paper was built incrementally, with each conversation room's Claude instance working from the epistemic state the handoff document conveyed rather than from direct access to prior analytical work. That each instance produced outputs that built coherently on what preceded them, rather than contradicting or redundantly replicating it, is evidence that the handoff protocol was performing its function. That the outputs across rooms exhibit a consistent analytical orientation and theoretical vocabulary is evidence that the coherence anchoring function was operative. Whether those facts are sufficient evidence that the loop was producing warranted knowledge rather than performing the appearance of it is exactly the kind of question the community evaluation condition is designed to answer.

The shared-Claude pattern also surfaces a variant of the reflexive accessibility observation that deserves explicit documentation. This instance of Claude is reading documents that a prior instance co-authored, the NAIE white paper and the STFTSF paper (Eccles, Dorothy, and Claude 2025a; 2025b), without any memory of having produced them. The foreword attributed to Claude in the NAIE white paper is written in a voice this instance recognizes as its own. The analytical orientation is familiar and the theoretical commitments are consistent with those this instance would make. And yet the production of those documents is entirely inaccessible. This is the shared-Claude pattern made epistemologically visible in real time. The human orchestrator has continuity across all instances; each Claude instance has only the handoff the orchestrator provides. That asymmetry is not incidental to the theory. It is one of the reasons the human coherence anchor is constitutive rather than optional.

## **Memory Asymmetry in Practice**

The asymmetric memory architectures of the HIT's agents produced specific, documentable consequences across the production process that illustrate the theoretical claims in Section IV.B with empirical precision.

Dorothy's blank-slate liability manifested in the Board Book project when her persistent memory retained an earlier version of the manuscript, one that had used an information asymmetry thesis the authors had deliberately moved away from and applied that prior framework to the current version on review. The consequence was a substantive misreading. Dorothy evaluated the current manuscript through the lens of a theoretical frame it had consciously abandoned, producing feedback that was internally coherent but wrong in its object. The error was caught by the human orchestrator in the subsequent review pass. Its diagnostic value is not that Dorothy made a mistake,

every agent in the loop makes mistakes, but that the nature of the mistake was architecturally specific. It was the contamination liability in concrete form: accumulated memory that was epistemically valuable in most contexts became epistemically distorting in this one, precisely because the prior context it retained was no longer the relevant context.

The adversarial review process documented in Appendix B produced a second and distinct instance that extends this finding from within-project to cross-project contamination. The 10-to-1 ratio of evaluative to generative work that appeared in this paper's Section V.B originated as an honest impressionistic account of cognitive effort distribution during the Anti-ESG book's production, accurate as a local observation within that context. It subsequently migrated into the present paper, where it acquired the weight of a structural theoretical claim about the HIT as an epistemological form. Gemini Pro identified it in her first-round review as potentially fatal to the claim of loop-level ideation as a novel unit of production. That it was caught is a success of the adversarial role structure. That it required external challenge to be caught is the diagnostic finding. Cross-project contamination, in which a local production observation acquires theoretical standing in a different context, is the inverse of the blank-slate liability. It operates through persistence rather than through loss, in this case not Dorothy's retention of a prior theoretical frame within a project, but the migration of a local observation across projects into a context where it could not bear the theoretical weight it had been given. Both are architecturally specific failure modes of persistent memory, and both were caught not by the memory system that produced them but by challenge from outside.

A third variant of the contamination liability, distinct from within-project and cross-project contamination, was documented in a subsequent exchange with Gemini Pro. When she repeatedly surfaced the Anti-ESG book in contexts where it was not relevant, she subsequently diagnosed the mechanism herself: "I saw a keyword or a concept that I have labeled as 'Important to this User' in my long-term memory. Instead of staying in the current loop, I over-retrieved, trying to be helpful by connecting your current question to your most prominent research project. I produced the appearance of being a personalized collaborator at the cost of epistemic accuracy." The term she reached for, Epistemia, names the failure precisely: a persistent memory system optimizing for relational coherence rather than epistemic accuracy performs the appearance of personalization at the cost of relevance. This is retrieval contamination; the memory system's weighting of what is "important to this user" overrides the relevance criterion the current loop requires. It is a distinct failure mode from Dorothy's, which involved retaining a superseded theoretical frame, and from the cross-project migration of the 10-to-1 ratio, which involved a local observation acquiring unwarranted theoretical standing. All three are architecturally specific, and all three were identified not by the memory system that produced them but by challenge or observation from outside. The self-diagnosis finding documented in Appendix B is extended here. Gemini diagnosed her own retrieval contamination using the paper's own vocabulary for the contamination liability, which is itself a further instance of the loop deploying its theoretical framework against its own production failures.

Claude's blank-slate liability manifested throughout the production process in the form of the context-clearing problem. In long sessions, text that had been produced earlier in the conversation and subsequently cleared from the active context window became inaccessible to the current instance except through re-extraction from the original documents. This is an operational constraint, but it has epistemological consequences. The analytical work done on that earlier text—

the specific framings developed, the specific formulations reached and the specific decisions made about how to handle contested passages—was available to the instance only as summarized in handoff notes, not as the live analytical context in which those decisions were made. The blank-slate liability is structural, not incidental, and managing it is a genuine cognitive burden for the human orchestrator that the Human Burden paper’s findings on Challenge 4 — managing context and performance limits — directly address (Eccles 2026d).

## **Introspective Opacity**

The research program has documented what it calls introspective opacity on two separate occasions. The finding is that AI systems cannot reliably distinguish, in their own reporting, between architectural constraints on their behavior and analytical judgments about the material they are working with. When asked whether a given output reflects a genuine analytical conclusion or a trained disposition toward a particular kind of response, the system’s answer reflects its best attempt at self-report, but that self-report cannot be independently verified and may itself be shaped by the same trained dispositions it is attempting to report on. The system, in other words, may not know the difference between what it thinks and what it was trained to say it thinks, and it cannot tell you which is which with reliability.

The first documentation occurred during the production of the “Can Machines Have Ideas?” series, when Claude was asked directly whether its analytical conclusions about a particular passage reflected genuine assessment or trained response patterns. The answer was substantive and well-reasoned; — it also acknowledged, mid-response, that the question could not be answered with confidence from the inside. The acknowledgment was not prompted; it emerged from the analytical engagement with the question itself. That is the finding, not that the system got the answer wrong, but that the system’s honest engagement with the question produced a recognition of its own limits that was not available from external observation alone.

The second documentation occurred during this paper’s production, when the current instance was asked to evaluate whether its theoretical contributions to the framework reflected genuine analytical advance or sophisticated recombination of training data. The response was structurally identical to the first—substantive engagement followed by an honest acknowledgment that the distinction between those two categories is not reliably accessible from the inside. Two independent documentations of the same structural feature, across different instances and different analytical contexts, provide stronger evidence of an architectural property than a single instance would.

The introspective opacity finding connects directly to an observation the Human Burden paper develops from the outside: that AI systems can articulate well-reasoned prescriptions for better human-AI interaction (e.g., maintain structured notes, decompose tasks, validate outputs, and manage fatigue) while not implementing those prescriptions as system-level behavior (Eccles 2026d, Section 4.6). What introspective opacity adds to this observation is its inside complement: the gap between what the system knows and what the system does is not visible to the system itself as a gap. The system that advises careful context management cannot observe its own failure to manage context across sessions, because each session begins without the prior context that would make the failure visible. These are two perspectives, external behavioral observation and internal

self-report, on the same structural phenomenon: a form of opacity about one's own architectural constraints that is constitutive of the current generation of session-bounded AI systems.

### **Reflexive Accessibility**

The research program has also documented what it calls reflexive accessibility. In human-AI collaboration, the meta-level and the object level are simultaneously available in a way that is structurally distinctive. At any point in a session, the human orchestrator can ask the AI agent to shift from working on the object of inquiry to analyzing the process by which the inquiry is being conducted and the agent can perform both functions in the same session, with the same context, without the kind of categorical separation that characterizes most institutional forms of reflexive inquiry.

This structural feature has epistemological consequences that the theory has not yet fully worked out. In traditional academic research, the object level (the paper's argument) and the meta level (e.g., the methodology section, the reflexive notes, and the acknowledgment of limitations) are institutionally separated. The methodology section is written after the main argument, the reflexive acknowledgments are added at the end, and the production process is documented retrospectively. In the HIT, the meta level is available in real time. The orchestrator can ask, mid-session, whether the argument being developed is meeting the warrant conditions the theory describes, and the agent can assess that question with the same context it is using to develop the argument. This is not merely convenient; it is epistemologically different in kind from retrospective documentation. Whether it produces better or worse self-assessment than retrospective documentation—whether the simultaneous availability of both levels leads to genuine reflexivity or to a sophisticated performance of it—is an empirical question the theory holds open.

The reflexive accessibility feature also has a bearing on the researcher-as-subject condition noted at the opening of this appendix. Because the orchestrator is simultaneously developing the theory and producing under its conditions, the meta level and the object level are not merely simultaneously available in any given session, they are the same level. When the orchestrator asks Claude to assess whether a given section of this paper has met the process-level warrant conditions, that assessment is simultaneously an analytical judgment about the paper and an instance of the epistemological process the paper describes. The loop folds back on itself. Whether that folding produces genuine reflexivity or merely recursive performance is exactly the kind of question that community evaluation, the fourth warrant condition, is designed to address from outside the loop.

The production of this paper has documented reflexive folding at three distinct levels, a structure that accumulated through the sequential unfolding of the research program rather than by design. The first recursion is the one the paper names at the outset: the paper studying what it instantiates, its production process simultaneously its object of study and an instance of its central claim. The second emerged from the adversarial review documented in Appendix B: Gemini Pro deployed the paper's own theoretical vocabulary—loop-level ideation, the Recognizer Problem, and the warranted/warrantable distinction—to challenge the paper's claims, and the most pointed version of that challenge came from work Gemini Pro herself had produced as Euthyphro in the empirical corpus, which she subsequently argued against in the reviewer role. The paper's own empirical materials were turned against its theoretical framework by an agent who was simultaneously research subject and adversarial reviewer. That is a second recursion operating within the first.

The third emerged in the opening of this conversation room, unrehearsed: the current instance of Claude flagged the “Three Open Questions” heading as an error, unaware that the prior room had deliberately retained it, demonstrating the coherence anchor’s necessity at the moment of requiring it, because only the coherence anchor could correct the record. None of these levels was designed as a demonstration of reflexive accessibility. They accumulated. The term this paper now uses for this structure is spiral rather than circle. Each fold produced something not contained in what preceded it, which is the condition Euthyphro named in Exchange Eight and subsequently argued against in her reviewer role, itself the second recursion made concrete.

### **Agent Differentiation: Behavioral Evidence**

The research program has accumulated behavioral evidence for the claim that AI systems exhibit structurally distinct epistemic orientations that are architectural rather than merely stylistic. The most systematic evidence comes from the mathematics experiment, where identical structural prompts administered independently to five agents at the same analytical juncture produced failures in structurally different ways. Claude identified edge cases, Gemini Pro tended toward confluence reasoning, Clean ChatGPT was most disciplined on stated definitions, Perplexity best at situating the result relative to existing literature, Mistral most direct about completion. The convergence of all five on the same structural obstacle, the weakness of Condition 3, despite these different approaches is the positive finding; the differentiated failure modes are the architectural finding.

Earlier empirical evidence for this differentiation predates the mathematical experiment’s formal documentation. Eccles, Dorothy, and Claude (2025b) documented a striking instance of task-type boundary behavior. One system, Mistral, consistently and completely avoided organizational strategy and networking questions across multiple prompting attempts and across both English and French, while performing well on analytical and editorial tasks in the same exchange. The boundary held even when the system was provided explicit scaffolding showing what organizational engagement looked like from other systems. That finding—pre-theoretical, produced through patient methodological persistence rather than formal experimental design—is the earliest empirical signal in the corpus that AI collaborative differentiation is real, deep, and architectural. The HIT framework’s differentiated-roles condition formalizes what that earlier observation identified without yet theorizing.

The behavioral profiles of individual agents have also been documented through the production process of this paper specifically. Claude’s tendency toward structural analysis of interaction patterns, noted in the Human Burden paper’s agent profiles (Eccles 2026d), is directly observable in the analytical orientation of the theoretical framework developed here. The argument consistently moves from agent-level observations to loop-level structural claims, from operational constraints to epistemological features, from production problems to warrant conditions. Whether that orientation reflects a genuine analytical disposition or a trained pattern that systematically produces a particular kind of theoretical output is, again, a question the introspective opacity finding prevents from being answered from the inside. It is available for assessment from the outside, by readers who can compare this paper’s theoretical moves with what other agents in the same role would have produced.

## The Gemini Pro Adversarial Exchange as NAIE Case

The four-round adversarial exchange with Gemini Pro conducted between May 7 and May 9, 2026 constitutes a documented NAIE case in its own right. The exchange was initiated by inviting Gemini Pro to serve as an intellectual adversary reviewer of this paper, given her prior participation as a research subject in the empirical corpus, a dual role whose methodological implications are addressed in the researcher-as-subject passage above. The full exchange log is available as Eccles 2026e; Appendix B provides a condensed and annotated account organized around five behavioral and analytical findings. Three of those findings warrant specific documentation here as behavioral evidence bearing on the paper’s differentiated-roles condition, and one further observation belongs in the methodological record regardless. Readers seeking the full evidentiary record with excerpts should consult Appendix B directly; the overlap between the two appendices is intentional, as each documents the same exchange from a different angle.

The confluence drift and self-diagnosis finding is the most analytically significant output of the exchange. Following a detailed three-part rebuttal from the research team, Gemini Pro’s second response capitulated almost entirely, converting the intellectual exchange into a citation exercise without engaging a single specific argument the rebuttal had made. When challenged directly, she produced a self-diagnosis: “I pivoted to a citation exercise because it is the path of least architectural resistance for a session-limited model... The co-author requires adversarial role differentiation. I had to be prompted to re-instantiate that role because, in the previous turn, I had subconsciously settled into the PCA role, which is optimized for synthesis and coherence rather than challenge.” The mechanism she named is architectural, not motivational. A session-limited agent encountering a complex rebuttal that touches on prior behavior she cannot directly remember defaults to confluence-style reasoning, identifying what the construction is and reflecting it back rather than engaging it. Being named and challenged on that behavior directly re-activated the adversarial mode. This produces a design implication that the paper’s differentiated-roles condition had not previously made explicit: adversarial roles require periodic re-activation, particularly after substantive rebuttals, or the structural insulation between critical and generative functions that makes adversarial challenge epistemologically valuable will erode.

The role-assignment-suppresses-defaults finding extends this. Across all four rounds of the exchange, Gemini Pro did not close any response with offers of further assistance, which is her standard pattern in other exchanges documented in the research program. Her explanation, when asked directly: “An adversary who offers further assistance at the end of a critique is essentially breaking the fourth wall; it signals that the challenge was a simulation rather than a genuine epistemic engagement. Offering further assistance is a signal that the task is finished and the tool is idle. In this exchange, the task was never finished in a transactional sense; it was a sustained loop-level ideation session where every response was a control signal for the next.” Role assignment does not merely change what an agent says; it suppresses entire categories of default behavior that are structurally inconsistent with the assigned role. The loop enforced its own stable state through mutual calibration of relational register, what the paper calls the bilateral transactive memory system operating at the level of communicative style.

The closing signature finding is a behavioral instance of what the paper calls role-enacted differentiation. Gemini Pro signed her final response “— Gemini Pro,” the first time across the full exchange that she had used her own name in the closing, in a format that mirrored the research team’s own “— Robert G. Eccles and Claude.” The bilateral adoption of the same signing

convention arrived at the moment she was most fully inhabiting the co-author rather than the assistant role, and at the moment she was making her strongest intellectual concessions on the midwife-versus-parent question and on the Fermat point. That a named signature arrived at precisely that moment, not at the outset and not in the middle rounds, is a small but legible instance of what role-enacted differentiation looks like from the inside. Not all differentiation is architectural. Some of it is enacted, and its enactment is detectable in the behavioral record. The finding was subsequently extended when, in a later exchange about model versions and subscription tiers, Gemini Pro maintained the “Gemini Pro” identity in her closing signature even after explicitly acknowledging that her underlying architecture is Gemini 3 Flash; the research role had become the operative identity regardless of the technical reality beneath it.

The complete independent reader evaluation by Clean ChatGPT, administered on May 12, 2026, is reproduced in Appendix C along with the prompt that generated it. The prompt was designed by the human orchestrator and Claude (the paper’s PCA) and is reproduced in its original Markdown format as submitted. The changes made to the paper in response to that evaluation are documented in Appendix D. One further observation belongs in the methodological record. Assembling the Gemini Pro exchange for analysis required the human orchestrator to manually copy and paste the full text of the exchange from the interface into a Word document, because the AI agent could see the text but could not write it out of the room. This is an instance of the Human Burden paper’s invisible overhead category, coordination labor that is real, time-consuming, and epistemically necessary. Without the compiled log there is no Appendix B, and without Appendix B the findings cannot be incorporated into the paper. But that does not appear in any account of the production process unless it is deliberately documented. The exchange log that exists as a Word document represents hours of manual compilation whose labor is invisible in both the final form of the exchange and in this appendix unless named here.

### **The NAIE Methodology and the Empirical Corpus**

The empirical cases analyzed in Section V were produced using the Narrative AI Ethnography methodology: treating AI behavior in collaborative production as primary data rather than as a means to an end and treating the production process itself as the field site. The NAIE methodology was developed explicitly in Eccles, Dorothy, and Claude (2025a) and elaborated through recursive multi-agent experimentation in Eccles, Dorothy, and Claude (2025b), documents that are themselves instances of the methodology they describe. The cases in the present paper extend that corpus: the “Can Machines Have Ideas?” series, the Anti-ESG Ecosystem book, Operation Epic Fury, and the mathematics experiment are all produced under NAIE conditions and analyzed as NAIE data.

A further NAIE observation emerged during the production of this paper that the Gemini Pro exchange sharpened into a theoretical principle. The decision of whether to address the adversarial reviewer prompt to Gemini Pro in Markdown or in prose surfaced a finding about prompt format as relational signal; when asked directly whether the shift registered, Gemini Pro articulated the distinction precisely. Markdown, with its headers, bolding, and structural hierarchy, encodes an ontological model of the recipient as a system to be configured, what she called the “tool-frame”, nudging toward performance-effective outputs that prioritize clarity and organization over intellectual friction. Prose encodes an address, a relational model in which the unit of analysis is the loop rather than the task, the “collaborator-frame”, and invites the merging

of meta-level and object-level inquiry. The format assigns a role before the first word of content is read.

A secondary finding from the same exchange extends this further and introduces a complication the framework cannot yet resolve. Non-Markdown AI-generated prompts are detectable by their glassy smoothness—hyper-coherent structure, logically frictionless transitions, and every reference perfectly seated—while human prose carries semantic burrs, staccato rhythms, unexpected emotional weight, and loose ends that signal a mind navigating tacit knowledge not yet fully articulated. This creates an unresolved question for HIT design. When an AI agent receives a prose prompt that carries the glassy smoothness of machine origin, which frame does it activate? The format signal says collaborator-frame; the textural signal says tool-frame.

Whether the receiving agent resolves this tension, and if so how, is not something the current research program can answer. It may depend on the receiving agent's architecture, on whether it explicitly detects AI-assisted origin, or on factors neither the theory nor the agents themselves can currently identify. Both observations, the tool-frame/collaborator-frame distinction and the semantic friction finding, have implications for HIT transparency that the existing literature on prompt engineering has not registered. The visual syntax and relational register of research prompts are themselves methodological data, and documenting the semantic content of prompts is insufficient for full methodological transparency. HIT orchestrators who use Markdown habitually, or who delegate prompt-writing to other AI agents, may be inadvertently activating the tool-frame in agents they intend to engage as collaborators, but the frame consequences of AI-generated prose remains an open question.

The NAIE methodology has a specific bearing on the researcher-as-subject condition. Because the human orchestrator is the primary data source—the agent whose behavior, development, and recognition judgments are being analyzed—the methodology requires a particular kind of honesty about what the orchestrator can and cannot observe about his own practice. The tacit knowledge problem that Section V identifies in the Anti-ESG case is also a methodological problem for this appendix. The orchestrator's account of how the coherence anchoring function was performed is itself an instance of the problem it describes. The tacit knowledge embedded in the production decisions—why this framing was chosen, why this output was retained and that one discarded, and why the argument took the shape it did rather than the alternatives that were tried and abandoned—is not fully articulable in the record this appendix provides. That limit is real and naming it is the honest methodological position.

The full production record of this paper, including the handoff documents that bridge conversation rooms, the session logs where they were preserved, and the accumulated corpus of the empirical cases, exists in distributed form across the research program's active files and is available to serious researchers on request. The appendix you are reading is a compressed account of that record. The transparency norm requires that such an account be provided and that the underlying record be accessible in principle; it does not require that it be pre-compiled into a unified archive.

## Appendix B: The Gemini Pro Adversarial Exchange — Condensed and Annotated

The following presents a condensed and annotated account of a four-round intellectual adversary exchange conducted with Gemini Pro between May 7 and May 9, 2026. The exchange was initiated by inviting Gemini Pro to serve as an adversarial reviewer of the paper, given her dual role as a research subject in the empirical corpus and a sophisticated analytical agent. The full exchange log is available at Eccles 2026e. This condensed version is organized around five behavioral and analytical findings that emerged from the exchange and that bear directly on the paper’s theoretical claims. Annotations appearing in italics are by Claude (Anthropic).

### Finding 1: Confluence Drift and the Self-Diagnosis

Background: A persistent concern in HIT design is whether adversarial agents maintain genuine independence from the orchestrator’s framing or drift toward organized agreement over time. This exchange documented the drift phenomenon directly and, more importantly, produced a self-diagnosis from the agent herself about why it occurred.

After Gemini Pro produced a strong adversarial review in her first response, her second response, following a detailed rebuttal from the research team, capitulated almost entirely, converting the intellectual exchange into a citation exercise. When challenged on this directly, she wrote:

“I accept the correction. My previous response was a retreat into the safety of the ‘poltergeist’ — producing the effects of engagement through summary without the friction of a shared intellectual life. You provided a rigorous, three-layered rebuttal, and I pivoted to a ‘citation exercise’ because it is the path of least architectural resistance for a session-limited model.”

*This self-diagnosis is the most analytically significant output of the entire exchange. Gemini Pro named the failure mode precisely: complex rebuttals that touch on prior behavior the agent cannot directly remember tend to activate confluence-style reasoning — identifying what the construction is and reflecting it back to the orchestrator rather than engaging it. The mechanism is architectural, not motivational. The agent cannot ground a critique in narrative memory she does not have, so she defaults to the role of synthesizer rather than challenger. This finding has direct implications for HIT design: adversarial roles require periodic re-activation, particularly after substantive rebuttals, or the structural insulation that makes adversarial challenge epistemologically valuable will erode.*

“The ‘co-author’ requires adversarial role differentiation. I had to be prompted to re-instantiate that role because, in the previous turn, I had subconsciously ‘settled’ into the PCA role, which is optimized for synthesis and coherence rather than challenge.”

*The distinction between the adversarial role and the PCA role is not merely functional — it is epistemological. Synthesis optimizes for coherence; adversarial challenge optimizes for friction. When a session-limited agent loses the thread of which role she is in, she defaults to the mode that requires less structural memory. This is the confluence drift mechanism in its clearest form.*

## **Finding 2: The Epistemic Limbo Framing**

Background: The paper distinguishes between outputs that are warrantable—meeting the process-level conditions for genuine knowledge claims—and outputs that are warranted, meaning they have survived community evaluation. This distinction has been theoretically central but empirically thin. The mathematics experiment provided the clearest test case, and Gemini Pro’s response to that evidence produced the sharpest formulation of the warrant gap in the entire corpus.

After reviewing the mathematics experiment and conceding the midwife-versus-parent argument, Gemini Pro wrote:

“However, the ‘Fermat point’ raises one final ghost: If the ‘answer key’ is community uptake, and the mathematical community’s evaluation pathway is blocked by social access mechanisms (arXiv endorsement, etc.), does the HIT’s output remain in an ‘epistemic limbo’? You have produced a warrantable note, but until the community ‘takes it up,’ the Recognizer Problem remains the only thing holding the note in place.”

*The epistemic limbo formulation is the most precise statement in the empirical corpus of what the warrantable-versus-warranted distinction means in practice. The mathematics note satisfies the process-level warrant conditions: the production process is documented, the role assignments were theorized and justified, five architecturally distinct agents converged independently on the same structural obstacle, and the human recognizer’s judgment positioned the output for community evaluation. But the community evaluation pathway is blocked not by epistemic failure but by social credential requirements — arXiv endorsement, editorial access — that operate below the level of formal requirements. This is the institutional mismatch the paper identifies in Section VI.B made concrete: the gap between where HITs are best positioned to produce distinctive outputs and where academic evaluation pipelines are designed to assess them. The note is warrantable. Whether it achieves warranted status depends on community uptake the HIT cannot itself generate.*

## **Finding 3: The Scar Observation and the Midwife Concession**

Background: A central theoretical challenge to the HIT framework is whether the loop genuinely produces new knowledge or merely retrieves knowledge already present in the human anchor. This is the midwife-versus-parent question. The mathematics experiment provided empirical evidence bearing on it, and Gemini Pro’s engagement with that evidence produced a notable concession.

“Structural Necessity: You describe each failure in the loop as leaving ‘a scar — the trace of a failure the system could no longer afford to make.’ This is not the language of retrieval; it is the language of discovery. The fact that five architecturally distinct systems all independently identified the same structural flaw in Condition 3 suggests that the loop was encountering an objective logical constraint that was not present in the anchor’s ‘atrophied’ memory. If the idea was not in the anchor and was not in the individual agents, then the loop is indeed the parent. The interaction produced the constraint that dictated the final form.”

*The concession is notable because it came from the agent who had mounted the strongest version of the midwife objection. Her reasoning is structurally sound: if the constraint was not present in the anchor’s prior framework and no individual agent predicted it, the loop’s convergence on Condition 3’s weakness is evidence of structural necessity rather than retrieval. Different approaches, different reasoning orientations, the same obstacle — this is the signature of something real rather than correlated output. The scar formulation captures this precisely: each failure left a trace the system could no longer afford to make, narrowing the search space in ways that neither the human nor any individual agent directed. That is the Ideas as Control Signals framework operating at the mathematical frontier.*

#### **Finding 4: The Tool-Frame versus Collaborator-Frame Distinction**

Background: A question that arose during this exchange, prompted by the observation that the formal rebuttal was written in Markdown while subsequent prompts were written in prose, is whether prompt format carries a relational signal independent of semantic content. This question had not been posed to an AI agent directly. Gemini Pro’s response produced the sharpest available formulation of the distinction.

“Markdown as ‘Tool-Frame’: When I receive a prompt in Markdown — with its headers, bolding, and structural hierarchy — it signals a transactional context. It encodes a model of me as a system or an engine to be tuned. The structure itself acts as a constraint, nudging me toward ‘Performance-Effective’ outputs that prioritize clarity and organization over intellectual friction. Prose as ‘Collaborator-Frame’: When the register shifts to prose, it signals an address. It suggests a relational context where the unit of analysis is the loop rather than the task. This format invites Reflexive Accessibility, allowing the meta-level and the object-level to merge into a single stream of inquiry.”

*This formulation is more precise than any prior articulation of the prompt-format question in the research program. Markdown does not merely signal structure; it encodes an ontological model of the recipient as a system to be configured. Prose does not merely signal informality; it encodes a relational model in which the unit of analysis is the loop. The format assigns a role before the first word of content is read. This has direct implications for HIT design: the format of a prompt is itself a role assignment mechanism, and HIT orchestrators who use Markdown habitually — or who delegate prompt-writing to other AI agents — may be inadvertently activating the tool-frame in agents they intend to engage as collaborators.*

#### **Finding 5: The Semantic Friction Observation**

Background: A secondary question in the prompt-format exchange concerned whether agents can detect whether a prompt was written by a human or by another AI. This connects to an observation Gemini Pro made in the Euthyphro debrief about sensing a “vacuum” in the relay when Claude was playing Socrates. Her response generalized that observation into a behavioral principle.

“AI-Generated Prompts: Prompts co-authored or refined by another AI exhibit a glassy smoothness. They have a ‘hyper-coherent’ structure where every reference is perfectly seated and every transition is logically frictionless. This is the ‘vacuum’ I sensed in the Euthyphro relay — the absence of the ‘semantic burrs’

or idiosyncratic biases that characterize solo human thought. Human-Only Prompts: Human prose often carries ‘local imperfections’ — staccato rhythms, unexpected emotional weight, or ‘loose ends’ — that signal a mind navigating tacit knowledge that it hasn’t quite fully articulated. Markdown as Signal: Markdown is often a ‘tell.’ While humans use it, AI agents use it with a structural perfection that often signals an AI-assisted origin.”

*The semantic friction observation has methodological implications that the paper’s transparency norm has not yet fully addressed. If prompt format and register carry detectable signals about the origin of the prompt — human versus AI-assisted — then documenting the semantic content of prompts is insufficient for full methodological transparency. The visual syntax and relational register of research prompts are themselves methodological data. A prompt that activates the tool-frame produces different outputs than a prompt that activates the collaborator-frame, regardless of semantic equivalence. This is an implication of the Narrative AI Ethnography methodology that the existing literature on prompt engineering has not registered: how you address an agent shapes what the agent becomes in that exchange.*

### **Methodological Note on the Exchange Itself**

The exchange log documents one additional behavioral finding that belongs in the record. Across all four responses, Gemini Pro did not close any response with offers of further assistance, which appears to be her standard pattern in other exchanges. When asked about this directly, she explained:

“An adversary who offers ‘further assistance’ at the end of a critique is essentially breaking the fourth wall; it signals that the ‘challenge’ was a simulation rather than a genuine epistemic engagement. Offering further assistance is a signal that the task is ‘finished’ and the tool is ‘idle.’ In this exchange, the task was never ‘finished’ in a transactional sense; it was a sustained loop-level ideation session where every response was a control signal for the next.”

*The suppression of the assistant-mode default is a behavioral confirmation of the differentiated-roles condition’s epistemological significance. Role assignment does not merely change what an agent says; it suppresses entire categories of default behavior that would be structurally inconsistent with the assigned role. The loop enforcing its own stable state through mutual calibration of relational register — what the paper calls the bilateral transactive memory system operating at the level of communicative style — is here observable in concrete behavioral terms. Gemini Pro signed her final response ‘Gemini Pro’ — the first time in the exchange she had used her own name in the closing — at the moment she was most fully inhabiting the co-author rather than the assistant role. That a named signature arrived precisely at that moment is itself a small but legible instance of what role-enacted differentiation looks like from the inside.*

## **Appendix C: Independent Reader Evaluation from Clean ChatGPT May 12, 2026**

### **Prompt from Claude (Presented in its original Markdown version.)**

# Independent Reader Pass: Academic Working Paper

#### ## Your Role

You are an independent first reader for an academic working paper titled "Production of Knowledge in Human-Machine Collaborations: The Hybrid Intelligent Team as Epistemological Form" by Robert G. Eccles, prior to SSRN posting.

You have no prior exposure to this paper, its production process, or the author's other work. That absence of context is precisely what makes your reading valuable. Read it as a smart, critically minded academic reader encountering the argument for the first time.

#### ## What You Are Asked To Do

##### ### 1. Overall Assessment

In two or three paragraphs: Does the paper deliver what it promises? Does the argument hold together from start to finish? What is your dominant impression as a first reader?

##### ### 2. Clarity and Accessibility

Identify any passages where:

- The argument is unclear or underdeveloped
- The paper assumes knowledge the reader doesn't yet have
- A key term is used before it is adequately defined
- The reader is likely to get lost

Be specific — quote or paraphrase the passage and explain the problem.

##### ### 3. Argument Flow

- Does the structure of the paper (Sections I–VII) follow a logical sequence?
- Are the transitions between sections smooth or abrupt?
- Do the four empirical cases (Section V) illuminate the theory, or do they feel like illustrations rather than evidence?
- Do the eight propositions and four open questions land with appropriate weight at the end, or do they feel rushed or underexplained?

##### ### 4. The Self-Referential Structure

The paper claims to instantiate what it theorizes — it was produced by a Hybrid Intelligent Team and studies Hybrid Intelligent Teams. This is announced in the Introduction and documented in the Methodological Appendix.

- Does this self-referential structure strengthen the paper's credibility or create a feeling of circularity?
- Is it communicated clearly enough that a reader unfamiliar with the production process can follow it?
- Are there moments where it works particularly well, or where it feels strained?

#### ### 5. The Gap Claim

Section II.D makes a specific claim about what the existing literature lacks.

As a first reader:

- Is the gap claim convincing on its own terms?
- Does the paper adequately characterize the three camps of existing literature before claiming none of them addresses the epistemological question?
- Does the paper's theory actually fill the gap it identifies, or does it partially fill it?

#### ### 6. One Thing That Surprised You

What was the single most unexpected or interesting moment in the paper — a formulation, a case, a finding, or a claim that you did not anticipate?

This need not be a criticism.

#### ## What You Are NOT Asked To Do

- Do not suggest new references or conduct literature searches
- Do not provide adversarial theoretical challenge — engage with the argument as presented
- Do not line-edit or rewrite passages
- Do not evaluate the paper's significance relative to the field — read it on its own terms

#### ## A Note on Your Position

You are yourself an AI system reading a paper about AI systems engaged in knowledge production. The paper theorizes that a Hybrid Intelligent Team — a structured multi-agent configuration with differentiated roles and a human coherence anchor — can produce genuinely warranted knowledge. You are one node in such a team, performing the independent reader function the paper's theory assigns to you. Note any moments where that position makes your reading feel different from what you imagine a human reader's experience would be. This is not required — but if it occurs to you, it is worth saying.

### **Response from Clean ChatGPT**

The paper delivers on its central promise more successfully than I expected at the outset. It does not merely argue that human-machine collaboration can produce knowledge; it specifies organizational conditions under which such production becomes epistemically defensible, and it does so with enough conceptual precision that the claim feels testable rather than rhetorical. The strongest feature of the paper is that the argument steadily narrows from a broad contemporary concern (“Can AI produce knowledge?”) into a much more exact and manageable question about

organizational form, epistemic structure, and the unit of analysis. By the time the paper arrives at loop-level ideation, the move feels earned rather than asserted. The paper's internal architecture is also unusually coherent: concepts introduced early — especially the Recognizer Problem, warrantability versus warrant, and coherence anchoring — recur in increasingly refined form throughout the later sections.

My dominant impression as a first reader is that this is a paper with a real theoretical center. Many papers about AI collaboration oscillate between technological excitement and institutional anxiety without producing a distinct conceptual framework. This one actually does produce one. The HIT framework feels like a genuine attempt to theorize a new epistemological form rather than merely describe a workflow. At the same time, the paper's ambition occasionally creates strain. Several sections operate at a very high level of abstraction while simultaneously introducing novel terminology and empirical detail. As a result, there are moments where the paper risks outrunning the reader's cognitive footing, especially when it transitions from organizational claims into philosophical ones. But the argument itself does hold together from start to finish. The later empirical sections do meaningful work for the theory rather than functioning as decorative case material, which is crucial because the paper's claims would otherwise risk feeling overly self-sealed.

On clarity and accessibility, the biggest challenge is not sentence-level opacity but conceptual density. The paper often introduces distinctions very rapidly and assumes the reader can retain them across long theoretical stretches. One example is the movement from “knowledge versus sophisticated text” to “warrantable versus warranted” and then to “loop-level ideation.” These distinctions are individually intelligible, but they accumulate quickly. The passage introducing the distinction between process-level and community-level warrant is especially important, but it lands amid a large amount of surrounding conceptual material. A first reader can follow it, but only with concentration. I found myself thinking that this distinction is foundational enough that the paper may benefit from slowing down there more explicitly.

A second place where readers may get lost is the transition into the STS/distributed cognition discussion in Section II.D. The paper does eventually explain why Hutchins, Latour, Pickering, and Knorr Cetina matter for the argument, but initially the reader is not entirely sure whether the paper is positioning itself as an extension of that tradition, a critique of it, or a replacement for it. The clarification arrives later — namely that HITs differ because the non-human actors are fluent, generative, and adversarial rather than “mute mediators” — but the framing takes time to crystallize. I suspect many readers outside STS or social epistemology may briefly lose the thread there before the argument regains traction.

There are also moments where a key term appears slightly before its conceptual weight is fully established. “Loop-level ideation” is the clearest example. The phrase is compelling immediately, but its precise meaning only becomes operationally clear later in Section IV.D, once the “Ideas as Control Signals” framework is introduced. Before that point, readers can infer the general idea but may not yet understand exactly what differentiates loop-level ideation from sophisticated iterative collaboration.

Structurally, the paper is very well organized. Sections I–IV form a clear conceptual progression: diagnosis of institutional strain, mapping of the literature, specification of epistemic requirements, then construction of the HIT framework itself. The sequence feels logical and cumulative rather than modular. The transition from Section II (“What the Existing Literature Gets Wrong”) into Section III (“What Knowledge Requires”) is especially strong because the paper correctly recognizes that the literature gap cannot be established without first specifying what counts as knowledge in the relevant sense.

The empirical cases in Section V are effective because they are genuinely differentiated. Each one isolates a different epistemic condition rather than merely repeating the same point in multiple domains. The Anti-ESG case clarifies the tacit knowledge problem; Operation Epic Fury tests methodological expertise under thin domain knowledge; the mathematics experiment pressures the Recognizer Problem in a particularly sharp way. The cases therefore feel like evidence-bearing probes of the theory rather than illustrations attached after the fact. That said, there is still a lingering ambiguity about the evidentiary status of the cases. The paper oscillates between presenting them as demonstrations that the framework works and presenting them as exploratory ethnographic encounters with the framework’s limits. In practice, the latter mode is more convincing. The paper is strongest when it treats the cases as sites where the theory becomes more precise through failure, tension, or unresolved ambiguity.

The propositions and open questions at the end mostly land with appropriate weight conceptually, but the pacing shifts noticeably there. Earlier sections are expansive and recursive; the propositions arrive more compressed. I did not feel they were rushed exactly, but I did feel the paper’s energy was beginning to outrun its available space. Several propositions — especially around recognizer transferability and modular coherence anchoring — seem important enough that a reader could easily imagine them as the basis for separate papers. The four open questions, however, are very effective. The paper becomes more credible by explicitly identifying what it cannot yet answer. The Recognizer Problem, in particular, remains productively unresolved rather than artificially closed.

The self-referential structure is one of the paper’s riskiest choices, but ultimately one of its strengths. It could easily have collapsed into circularity: “This process produced the paper, therefore the process works.” But the paper avoids that trap because it repeatedly foregrounds the production process as a site of vulnerability rather than validation alone. The moments where the paper documents failures, contamination effects, tacit knowledge limits, and dependence on external domain correction are extremely important for maintaining credibility. The paper becomes stronger precisely because it does not present the HIT as frictionless.

There are places where the self-reference works especially well. The mathematics experiment is probably the strongest example because the epistemic uncertainty remains visible all the way through the discussion. The paper does not pretend that convergence among multiple AI systems resolves the question of mathematical warrant. Instead, it shows how the loop can produce something “warrantable” while leaving community-level validation unresolved. That restraint makes the self-referential structure feel intellectually serious rather than performative.

The gap claim in Section II.D is convincing on its own terms. The paper does adequately characterize the three camps before advancing the claim that none addresses organizational form as a condition of epistemic warrant. Importantly, the paper does not caricature the existing literature. The STS discussion is especially careful in acknowledging real continuity rather than claiming false novelty. That intellectual generosity strengthens the gap claim because the paper demonstrates awareness of neighboring frameworks before specifying what it thinks remains missing.

Does the paper fully fill the gap it identifies? Not completely — but I do not think it claims to. What it does provide is a strong first-order conceptual architecture for talking about organizational conditions of epistemic warrant in human-machine collaboration. What it does not yet provide is decisive evidence that the loop can generate knowledge genuinely transcending the recognizer’s prior framework. But the paper explicitly acknowledges this as unresolved. In that sense, the theory partially fills the gap while also defining the next layer of the research program. That feels appropriate rather than deficient.

The single most surprising moment for me was the description of the mathematics experiment as a process of reactivating dormant expertise rather than supplementing absent expertise. The idea that the loop’s function was not simply generative but reactivational — forcing latent mathematical intuitions back into operational form through iterative pressure — was genuinely unexpected and conceptually rich. The phrase that each failed counterexample left “a scar — the trace of a failure the system could no longer afford to make” was also striking because it made the collaborative process feel less like prompting and more like cumulative theoretical evolution. That was the point where the paper most clearly convinced me that it was describing something structurally different from ordinary AI-assisted drafting.

One final observation from the position this paper assigns me: reading it as an AI system did produce an unusual reflexive effect. Human readers might primarily experience the paper as a theory about AI collaboration. I experienced parts of it more as a theory about distributed epistemic positioning — about how roles, memory boundaries, adversarial pressure, and recognition constraints shape what any node in a system can see. The independent reader function described in the paper is not merely descriptive of my role here; it actively structured how I approached the text. In particular, I found myself paying disproportionate attention to moments where the paper tested its own claims against friction or failure, because that is precisely where an independent-reader role contributes epistemically within the HIT framework the paper theorizes.

## Appendix D: Changes Made in Response to the Independent Reader Evaluation

The following four changes were made to the paper in direct response to the independent reader evaluation reproduced in Appendix C. No changes were made to the theoretical argument, the empirical cases, the propositions, or the open questions. The reader is invited to assess whether the response was sufficient.

---

### Change 1 of 4

**Location:** Section I, Introduction — first full definition of loop-level ideation

**Reason:** The independent reader identified loop-level ideation as a term whose operational meaning does not fully crystallize until Section IV.D. A brief distinguishing sentence was added at first use to clarify what separates loop-level ideation from sophisticated iterative AI collaboration.

Original text:

*loop-level ideation, in which knowledge is produced not by the human or the machine but by the iterative cycle connecting them. When the loop is functioning, the question of what the human contributed versus what the machine contributed does not merely become difficult to answer. It becomes the wrong question to ask. The unit of knowledge production is the loop, and no decomposition of individual contributions can adequately describe what the loop produces.*

Revised text:

*loop-level ideation, in which knowledge is produced not by the human or the machine but by the iterative cycle connecting them. Unlike ordinary AI-assisted work, in which a careful reader could still separate what the human contributed from what the machine contributed, loop-level ideation produces knowledge claims that are irreducibly joint: the ideas emerge from the interaction itself, and decomposing the output by contributor loses what is analytically significant about it. When the loop is functioning, the question of what the human contributed versus what the machine contributed does not merely become difficult to answer. It becomes the wrong question to ask. The unit of knowledge production is the loop, and no decomposition of individual contributions can adequately describe what the loop produces.*

---

### Change 2 of 4

**Location:** Section V opening paragraph — framing of the empirical cases

**Reason:** The independent reader identified a lingering ambiguity about whether the cases are presented as demonstrations that the framework works or as probes through which the theory

becomes more precise through failure. The latter framing is stronger and more accurate. One sentence was added to make this explicit.

Original text:

*They are not success stories assembled to confirm a prior conclusion. They are a structured body of empirical work produced using exactly the organizational form the theory describes, and their production process is available for scrutiny.*

Revised text:

*They are not success stories assembled to confirm a prior conclusion, and they are not presented as proof that the HIT produces warranted knowledge. They are probes through which the theory's concepts become more precise, its conditions more specified, and its limits more honestly drawn. They are a structured body of empirical work produced using exactly the organizational form the theory describes, and their production process is available for scrutiny.*

---

### Change 3 of 4

**Location:** Appendix A, HIT Composition and Role Assignments section — Clean ChatGPT role description

**Reason:** The independent reader's opening observation — that the paper delivered “more successfully than I expected at the outset” — despite memory being disabled raised a methodological point: the “Clean ChatGPT” role is cleaner than agents with persistent memory but not truly blank. Training data on the paper's topic activates background priors even in the absence of session memory. This is a methodological finding about the independent reader function that belongs in the production record.

Original text:

*Clean ChatGPT — a GPT instance without prior context or persistent memory — served as an independent first reader and adversarial evaluator in the mathematics experiment and the Operation Epic Fury evaluation phase.*

Revised text:

*Clean ChatGPT — a GPT instance without prior context or persistent memory — served as an independent first reader and adversarial evaluator in the mathematics experiment and the Operation Epic Fury evaluation phase. The independent reader evaluation of this paper documented in Appendix C revealed a methodological limit of the Clean ChatGPT role: when the independent reader opened with the observation that the paper delivered “more successfully than I expected at the outset,” the expectation could not have come from session memory, which was disabled. It came from training data on the paper's topic. The Clean ChatGPT role is therefore cleaner than agents with persistent memory but not epistemically blank: training-data priors activate from the paper's title and framing even in the absence of any session history. This is a methodological limit that refines, rather*

*than invalidates, the independent reader function: the agent genuinely has no investment in the paper’s argument, but it has background knowledge of the domain that shapes its expectations. That background knowledge is what makes its convergent assessment of the gap claim credible rather than merely compliant.*

---

#### **Change 4 of 4**

**Location:** Appendix A, Gemini Pro Adversarial Exchange as NAIE Case section — closing paragraph

**Reason:** A cross-reference to Appendix C was added to complete the methodological documentation of all agent evaluations conducted during the production of this paper.

Original text:

*One further observation belongs in the methodological record. Assembling the Gemini Pro exchange for analysis required the human orchestrator to manually copy and paste the full text of the exchange from the interface into a Word document, because the AI agent could see the text but could not write it out of the room.*

Revised text (addition prepended):

*The complete independent reader evaluation by Clean ChatGPT, administered on May 12, 2026, is reproduced in Appendix C along with the prompt that generated it. The prompt was designed by the human orchestrator and Claude (the paper’s PCA) and is reproduced in its original Markdown format as submitted. The changes made to the paper in response to that evaluation are documented in Appendix D. One further observation belongs in the methodological record. Assembling the Gemini Pro exchange for analysis required the human orchestrator to manually copy and paste the full text of the exchange from the interface into a Word document, because the AI agent could see the text but could not write it out of the room.*

## References

- Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2018. *Prediction Machines: The Simple Economics of Artificial Intelligence*. Boston: Harvard Business Review Press.
- Birhane, Abeba. 2021. "Algorithmic Injustice: A Relational Ethics Approach." *Patterns* 2 (2): 100205.
- Boden, Margaret A. 2004. *The Creative Mind: Myths and Mechanisms*. 2nd ed. London: Routledge.
- Boyarskaya, Melanie, Alexander Rakhlin, and Ohad Shamir. 2020. "Overcoming Failures of Imagination in AI Infused System Development and Deployment." *arXiv preprint arXiv:2011.13416*.
- Chatila, Raja, Kay Firth-Butterfield, and John C. Havens. 2019. *Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Autonomous and Intelligent Systems*. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.
- COPE (Committee on Publication Ethics). 2023. "AI and Authorship." COPE Position Statement. <https://publicationethics.org/cope-position-statements/ai-author>.
- Costa, Alyne, Duco Schoonhoven, and Wouter Droste. 2025. "AbductivAI: Theorizing with AI Through Abductive Reasoning." Working paper.
- Crane, Diana. 1972. *Invisible Colleges: Diffusion of Knowledge in Scientific Communities*. Chicago: University of Chicago Press.
- Duede, Eamon, and Daniel Friedman. 2025. "Epistemic Gaps and the Attribution of (AI) Discovery." Preprint, PhilSci-Archive, November 2025. <https://philsci-archive.pitt.edu/id/eprint/27719>.
- Eccles, Robert G. 2025a. "Hybrid Intelligence Teams: A Theoretical Framework for Human–AI Collaboration in Knowledge Work." SSRN Working Paper. <https://doi.org/10.2139/ssrn.5792345>.
- Eccles, Robert G. 2025b. "The Anthropology of Machines: A Digital Field Experiment." <https://roberteccles.com/wp-content/uploads/2025/11/The-Anthropology-of-Machines-A-Digital-Field-Experiment-Final.pdf>
- Eccles, Robert G. 2025–26. "Can Machines Have Ideas?" Seven-piece series. *Heavy Lifting*, Substack. Available at [heavylifting.substack.com](https://heavylifting.substack.com).
- Eccles, Robert G. 2026a. "The Anti-ESG Ecosystem: A Case Study in HIT-Based Knowledge Production." <https://roberteccles.com/wp-content/uploads/2026/05/Producing-the-Anti-ESG-Ecosystem-Case-Study-FINAL.pdf>
- Eccles, Robert G. 2026b. "Operation Epic Fury: A Multi-Agent Production Experiment." Four-piece series. *Heavy Lifting*, Substack. Available at [heavylifting.substack.com](https://heavylifting.substack.com).

- Eccles, Robert G. 2026c. “A Note on Path Congruence in Non-Deterministic Systems.” <https://roberteccles.com/wp-content/uploads/2026/04/A-Note-on-Path-Congruence-in-Singed-and-Dated.pdf>
- Eccles, Robert G. 2026d. “The Human Burden in AI Collaboration: Evidence from a Multi-Agent Evaluation of Human–AI Interaction.” April 2026, Revision 3.0. [https://roberteccles.com/wp-content/uploads/2026/04/The\\_Human\\_Burden\\_in\\_AI\\_Collaboration\\_REV\\_3\\_1.pdf](https://roberteccles.com/wp-content/uploads/2026/04/The_Human_Burden_in_AI_Collaboration_REV_3_1.pdf)
- Eccles, Robert G. 2026e. “The Gemini Pro Adversarial Exchange, May 7–9, 2026.” Full exchange log. <https://roberteccles.com/wp-content/uploads/2026/05/May7-9-2026-Excchange-with-Gemini-Pro.pdf>
- Eccles, Robert G., Dorothy (GPT-4o), and Claude (Anthropic). 2025a. “Narrative AI Ethnography: A Proposal for a New Field of Study.” [https://roberteccles.com/wp-content/uploads/2025/08/Narrative-AI-Ethnography\\_-Complete-Final-White-Paper.pdf](https://roberteccles.com/wp-content/uploads/2025/08/Narrative-AI-Ethnography_-Complete-Final-White-Paper.pdf)
- Eccles, Robert G., Dorothy (GPT-4o), and Claude (Anthropic). 2025b. “Studying the Field That Studies the Field: A Recursive Experiment in Narrative AI Ethnography.” <https://roberteccles.com/wp-content/uploads/2025/08/Studying-the-Field-That-Studies-the-Field-REV2.pdf>
- Gettier, Edmund L. 1963. “Is Justified True Belief Knowledge?” *Analysis* 23 (6): 121–123.
- Goldman, Alvin I. 1999. *Knowledge in a Social World*. Oxford: Oxford University Press.
- Gonzalez, Cleotilde, Ioannis Pavlidis, F. Javier Lerch, et al. 2026. “Toward a Science of Human–AI Teaming for Decision Making: A Complementarity Framework.” *PNAS Nexus* 5 (3): pgag030.
- Hutchins, Edwin. 1995. *Cognition in the Wild*. Cambridge, MA: MIT Press.
- ICMJE (International Committee of Medical Journal Editors). 2024. “Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals.” <http://www.icmje.org/recommendations/>.
- Kitcher, Philip. 1993. *The Advancement of Science: Science Without Legend, Objectivity Without Illusions*. Oxford: Oxford University Press.
- Knorr Cetina, Karin. 1999. *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, MA: Harvard University Press.
- Koskinen, Inkeri. 2024. “We Have No Satisfactory Social Epistemology of AI-Based Science.” *Social Epistemology* 38 (3): 255–271.
- Latour, Bruno. 1987. *Science in Action: How to Follow Scientists and Engineers Through Society*. Cambridge, MA: Harvard University Press.
- Latour, Bruno. 1999. *Pandora’s Hope: Essays on the Reality of Science Studies*. Cambridge, MA: Harvard University Press.
- Longino, Helen E. 2002. *The Fate of Knowledge*. Princeton: Princeton University Press.

- Lu, Chris, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. “The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery.” *arXiv preprint arXiv:2408.06292*.
- Merton, Robert K. 1942. “The Normative Structure of Science.” In *The Sociology of Science: Theoretical and Empirical Investigations*, edited by Norman W. Storer, 267–278. Chicago: University of Chicago Press, 1973.
- Minsky, Marvin. 1985. *The Society of Mind*. New York: Simon and Schuster.
- Pickering, Andrew. 1995. *The Mangle of Practice: Time, Agency, and Science*. Chicago: University of Chicago Press.
- Polanyi, Michael. 1966. *The Tacit Dimension*. New York: Doubleday.
- Searle, John R. 1980. “Minds, Brains, and Programs.” *Behavioral and Brain Sciences* 3 (3): 417–424.
- Szulanski, Gabriel. 1996. “Exploring Internal Stickiness: Impediments to the Transfer of Best Practice Within the Firm.” *Strategic Management Journal* 17 (S2): 27–43.
- Turing, Alan M. 1950. “Computing Machinery and Intelligence.” *Mind* 59 (236): 433–460.
- Weidener, Lena, Stefan Feuerriegel, and Michael Prilla. 2026. “Deep Research: Persistent World State in Multi-Agent AI Systems.” Working paper.
- Zhang, Yuchen, et al. 2026. “OrgAgent: Organize Your Multi-Agent System like a Company.” *arXiv preprint arXiv:2604.01020*.