Hybrid Intelligence Teams: A Theoretical Framework for Human–AI Collaboration in Knowledge Work

ABSTRACT

This paper develops a theory of **Hybrid Intelligence Teams (HITs)**—persistent groups composed of multiple humans and multiple AI agents working interdependently to accomplish knowledge-intensive tasks. We argue that traditional models of human teams, human–automation systems, and multi-agent AI are insufficient for explaining the emergent dynamics of these hybrid systems. Building on organizational behavior, human–AI interaction, distributed cognition, and hybrid intelligence research, we extend the IMOI (Inputs–Mediators–Outputs–Inputs) model to include hybrid constructs: **cross-species shared mental models**, **bilateral transactive memory systems**, **epistemic safety**, and **coherence anchoring**. We further identify hybrid-specific failure modes—such as authority confusion, information cascades, AI consensus illusions, and cross-agent drift—that do not appear in purely human or purely artificial teams.

We propose a unified HIT framework integrating:

(1) hybrid composition,
(2) hybrid mediators,
(3) hybrid interaction processes, and
(4) distinctive hybrid outputs.

The paper concludes with a research agenda for empirical study of HIT performance, design, measurement, and governance.

Methodological note:
All prose in this paper was generated by AI systems (ChatGPT, Claude, and Perplexity) under human supervision. **Appendix A** documents the workflow for methodological transparency. The theoretical contribution, construct architecture, and iterative direction were provided by the human researcher; all text was AI-generated.

KEYWORDS: Hybrid Intelligence Teams; Human–AI Collaboration; Multi-Agent Systems; Shared Mental Models; Transactive Memory; Distributed Cognition; Organizational Behavior; Generative AI

I. INTRODUCTION: FROM REPLACEMENT TO AMPLIFICATION

For more than a decade, public discourse surrounding artificial intelligence has been dominated by a single question: Will AI replace human workers? This framing, intuitive and emotionally compelling as it may be, increasingly appears inadequate for explaining the most important effects of generative AI in knowledge work. The question itself inherits a mental model from industrial automation—where machines substituted for human physical labor in manufacturing, agriculture, and transportation—and projects it onto the domain of cognitive work. Yet generative AI represents a categorically different phenomenon. Unlike previous waves of automation that displaced routine manual tasks, large language models and other generative systems augment, extend, and amplify human reasoning rather than merely substituting for it.

The shift from replacement to amplification as a dominant—though not exclusive—paradigm represents what may prove to be one of the most consequential conceptual pivots in twenty-first-century theories of work, even as displacement concerns persist in many sectors. This reframing compels us to revisit nearly every assumption about expertise, skill development, team composition, and knowledge creation in organizational settings. As Brynjolfsson and McAfee (2014) observed in their analysis of the second machine age, technological change is not merely about substitution but about complementarity—and nowhere is this complementarity more pronounced than in the interaction between human experts and advanced AI systems. What we are witnessing is not human cognition being rendered obsolete, but rather the emergence of a new unit of analysis for understanding productive knowledge work: the human–AI collaborative dyad, and more significantly, the multi-human, multi-AI team.

The Fundamental Reframing: Competition vs. Collaboration

When we pose the question 'Will AI replace humans?' we implicitly construct a competitive framework with two discrete categories: human workers and machine systems. This binary framing made sense in the context of industrial automation, where robotic assembly lines physically displaced human assembly workers, or where automated telephone systems substituted for human operators. In those contexts, substitution was indeed zero-sum: each robot installed on a factory floor directly corresponded to a reduction in human labor requirements.

Generative AI, however, operates according to fundamentally different principles. These systems do not perform discrete, bounded tasks in isolation from human oversight. Rather, they engage in open-ended cognitive processes—reasoning, ideation, synthesis, critique, pattern recognition—that interweave with human judgment at multiple stages of knowledge work. As Autor (2015) demonstrated in his analysis of task-based technical change, the crucial distinction is between routine and non-routine tasks, with AI complementing rather than replacing human expertise in complex, judgment-intensive domains. The cognitive labor performed by large language models is inherently collaborative: models generate hypotheses that humans evaluate; humans provide framings that models elaborate; models surface patterns that humans contextualize; humans articulate intuitions that models formalize.

Harvard Business School's work on AI and organizational strategy has emphasized this transformation, noting that competitive advantage increasingly accrues not to organizations with AI or to organizations with human talent, but to organizations that effectively integrate both (Iansiti & Lakhani, 2020). This formulation reorients the competitive dynamic entirely. The relevant comparison is no longer between human and machine performance, but between augmented and unaugmented human performance. The locus of competition shifts from human vs. machine to human+machine vs. human, and the strategic question facing organizations becomes not whether to deploy AI, but how to integrate AI into workflows in ways that maximize human–machine complementarity.

This distinction has profound implications. It suggests that the primary impact of generative AI in knowledge work will not be mass unemployment, but rather a dramatic widening of productivity distributions within knowledge sectors. Those who learn to collaborate effectively with AI systems may experience substantial gains in capability; those who resist or fail to develop collaborative skills will find themselves increasingly disadvantaged. The technology thus acts as a skill-complementary innovation (Acemoglu & Restrepo, 2018), amplifying returns to expertise rather than rendering expertise obsolete.

Amplification as a Skill-Complementary Technology

Empirical research across multiple domains—software development, scientific writing, strategic analysis, creative design, and medical diagnosis—has begun to reveal a consistent pattern: AI tools raise average performance across populations, but they raise top-tier performance even more dramatically, while providing more modest benefits to those lacking foundational domain expertise. This is the signature of a skill-complementary technology, one where productivity gains accrue disproportionately to those already possessing high levels of human capital.

A landmark study by Noy and Zhang (2023) examining the impact of ChatGPT on professional writing tasks found that while AI assistance improved performance across the distribution, substantial gains accrued to writers who were already high performers. Similarly, research by Peng et al. (2023) on GitHub Copilot demonstrated that experienced programmers integrated AI-generated code suggestions far more effectively than novices, who often accepted erroneous suggestions without adequate vetting. In both cases, domain expertise proved essential for extracting value from AI collaboration.

Why does AI amplification favor existing experts? The mechanism appears to be rooted in the complementarity between human pattern recognition and AI's generative capacity. Experts possess rich mental models of their domains—they know which questions matter, which patterns are meaningful, what constitutes a plausible answer, and where errors typically occur. When experts collaborate with AI systems, they can leverage the system's computational power and broad knowledge base while simultaneously filtering outputs through their own refined judgment. They recognize when AI-generated content requires correction, when it offers genuine insight, and how to integrate machine-generated ideas into coherent frameworks.

Non-experts, by contrast, lack the cognitive scaffolding necessary to evaluate AI outputs effectively. They cannot easily distinguish between plausible-sounding but incorrect statements

and genuinely accurate information. They struggle to integrate AI-generated fragments into coherent wholes. They miss opportunities to steer AI systems toward more productive lines of inquiry through skilled prompting. As Dell'Acqua et al. (2023) demonstrated in their study of consultants using GPT-4, the AI's impact was highly dependent on the user's ability to critically evaluate and iteratively refine outputs—a capability that correlated strongly with prior domain expertise.

The implications are significant. In elite fields—academic research, strategic consulting, advanced engineering, financial analysis, executive decision-making—the difference between good practitioners and exceptional ones is already substantial. Exceptional practitioners do not simply know more; they think differently. They construct richer mental models, generate more sophisticated hypotheses, integrate information more holistically, and evaluate evidence more rigorously. When such practitioners gain access to AI systems that can generate alternative framings, surface non-obvious patterns, provide instant access to vast literatures, and offer real-time critique of reasoning processes, their advantage may compound substantially.

The Second-Order Effect: Idea Generation Outpaces Execution Capacity

An intriguing second-order consequence of AI amplification is now emerging in some knowledge-intensive contexts: experts report generating conceptual possibilities faster than they can execute or evaluate them. This represents a potential inversion of traditional bottlenecks in intellectual work. Historically, the scarce resource in research, strategy formulation, and complex problem-solving was cognitive capacity—the ability to generate novel ideas, identify relevant patterns, or synthesize disparate information. Execution, while demanding, was downstream from ideation.

Generative AI may disrupt this hierarchy in certain contexts. When experts can delegate significant portions of ideation, synthesis, and preliminary analysis to AI systems, they may find themselves managing an abundance of possibilities. The constraint shifts from generating ideas to curating, evaluating, prioritizing, testing, and integrating them into coherent frameworks. This observation aligns with theories of bounded rationality and attention as a scarce resource (Simon, 1971), though systematic empirical investigation is required to establish the scope and magnitude of this phenomenon.

This shift has several important conjectures for how knowledge work is organized. First, it may elevate the importance of evaluative judgment. In an environment where generating multiple potential solutions to a problem takes minutes rather than days, the ability to rapidly assess which solutions are worth pursuing becomes paramount. This places a premium on metacognitive skills—the capacity to think about thinking, to recognize when an idea is genuinely promising versus merely novel, and to make fast, accurate judgments about intellectual quality.

Second, it may intensify the need for team-level coordination. When individual contributors can each generate vast arrays of possibilities, the challenge of integrating those possibilities into unified organizational action becomes more acute. Teams must develop mechanisms for filtering, synthesizing, and converging on shared priorities—mechanisms that were less critical

when ideation itself was rate-limiting. This observation points toward the central argument of this paper: as AI amplifies individual cognitive capacity, the bottleneck may shift to team-level cognition and coordination.

From Individual Amplification to Team Transformation

If the first-order effect of generative AI is individual amplification, the second-order effect—still largely untheorized—is team transformation. Nearly all consequential knowledge work occurs not through isolated individuals but through teams: research collaborations, product development groups, strategic planning committees, creative partnerships, and executive leadership teams. Yet the literature on AI and work has focused almost exclusively on dyadic human–AI interaction, treating team-level dynamics as secondary.

This oversight is consequential. Teams are not simply collections of individuals; they are cognitive systems in their own right, with emergent properties that cannot be reduced to individual capabilities. As Woolley et al. (2010) demonstrated, teams exhibit a 'collective intelligence' that predicts performance above and beyond the intelligence of individual members. This collective intelligence arises from patterns of interaction, communication norms, shared mental models, and coordination mechanisms. The concept resonates with distributed cognition theory (Hutchins, 1995), which treats cognitive processing as distributed across individuals, artifacts, and environmental structures, and with extended mind philosophy (Clark & Chalmers, 1998), which argues that cognitive systems can span brain, body, and world.

When we introduce AI into teams—and especially when we introduce multiple AI agents with different capabilities—we are not merely augmenting individual team members. We are fundamentally restructuring the team's cognitive architecture. The team becomes a hybrid system: partly human, partly artificial, with reasoning distributed across both biological and computational substrates. This hybrid configuration raises profound questions that existing literature has not adequately addressed.

How should roles be allocated between humans and AI agents? How do teams maintain shared mental models when some team members are human and others are AI systems with fundamentally different cognitive architectures? What communication protocols enable effective coordination in hybrid teams? How should teams manage the integration of divergent outputs from multiple AI agents? What leadership practices are required to maintain coherence when cognitive contributions come from both human and artificial sources?

These questions define an emerging frontier of research on AI and work. They point toward the need for a comprehensive theory of hybrid intelligence teams—teams that integrate multiple humans and multiple AI agents into unified cognitive systems. Such a theory must synthesize insights from organizational behavior, human–computer interaction, multi-agent systems, cognitive science, and sociotechnical systems design. The remainder of this paper develops such a theory, building on decades of research on human team effectiveness while extending that research to accommodate the novel challenges posed by human–AI collaboration.

II. THEORETICAL FOUNDATIONS: WHAT ORGANIZATIONAL BEHAVIOR KNOWS ABOUT HIGH-PERFORMING TEAMS

Before proposing a novel theory of hybrid intelligence teams, we must first establish what organizational behavior already knows about team effectiveness. Six decades of research have produced robust insights into the mechanisms that enable human teams to coordinate effectively, integrate diverse expertise, and produce outcomes superior to what individuals could achieve alone. These insights provide the foundation upon which any theory of human–AI teams must build.

The core proposition is this: hybrid teams will not succeed merely because they incorporate AI capabilities. They will succeed to the extent that they instantiate the same fundamental coordination and integration mechanisms that make all-human teams effective—while also addressing new challenges unique to human–AI configurations. Put differently, AI changes the implementation details of team effectiveness but not its underlying principles. We must understand those principles before we can extend them.

Shared Mental Models: The Foundation of Implicit Coordination

Perhaps the most extensively studied mechanism underlying team effectiveness is the shared mental model (SMM). As Cannon-Bowers, Salas, and Converse (1993) argued in their foundational work, team members develop shared cognitive representations of four key domains: the task itself, the equipment and tools available, the situation or problem context, and the team's structure and member roles. When team members possess overlapping mental models in these domains, they can anticipate each other's needs, coordinate implicitly without constant communication, and adapt fluidly to changing circumstances.

The importance of shared mental models has been demonstrated across domains ranging from aviation crews to surgical teams to software development teams (Espinosa et al., 2007). In each context, teams with greater SMM accuracy and overlap outperform teams whose members possess divergent or incomplete mental models. The mechanism is straightforward: shared mental models reduce coordination costs, minimize misunderstandings, enable rapid adaptation to novel situations, and support the development of team routines that become increasingly efficient over time.

What implications does this have for hybrid teams? Most directly, it suggests that effective human–AI teams will require shared mental models that bridge human and artificial team members. Humans will need accurate models of what AI agents can do, how they reason, where they excel, and where they fail. Conversely, AI systems will need representations of human priorities, decision criteria, risk preferences, and domain knowledge—essentially, models of their human teammates' expertise and judgment.

This creates what we might call cross-species shared mental models—cognitive representations that span fundamentally different types of intelligence. The challenge is non-trivial. Humans

develop mental models of other humans through interaction, observation, and social learning, leveraging their folk psychological theories about belief, desire, and intention. But AI agents do not have beliefs or desires in any conventional sense, and their reasoning processes are often opaque even to their designers. Similarly, current AI systems lack the sophisticated social cognition necessary to build rich models of individual human team members.

Nevertheless, approximate shared mental models remain achievable and valuable. Humans can develop heuristic models of AI capabilities through repeated interaction: learning which AI agent produces the most creative outputs, which is most reliable for fact-checking, which tends toward verbose explanations versus concise summaries. AI systems can be fine-tuned or prompted to reflect human preferences, domain norms, and quality standards. While these models will never achieve the richness of human social understanding, they can provide sufficient overlap to enable effective coordination.

Transactive Memory Systems: Distributed Knowledge Architecture

If shared mental models enable teams to coordinate on tasks and roles, transactive memory systems (TMS) enable them to coordinate on knowledge itself. Introduced by Wegner (1987) and elaborated extensively by Moreland and colleagues (Moreland & Myaskovsky, 2000; Lewis, 2003), TMS refers to the meta-knowledge structure through which team members keep track of who knows what. Rather than each team member maintaining comprehensive knowledge across all domains, team members specialize—and crucially, they maintain a shared directory of those specializations.

The efficiency gains from TMS are substantial. Teams with well-developed transactive memory can access specialized knowledge quickly by directing queries to the appropriate expert. They avoid redundancy, as members do not duplicate each other's learning efforts. They adapt more readily to personnel changes, as the meta-knowledge structure helps new members quickly identify where to seek information. And they leverage cognitive diversity more effectively, as specialization allows depth of expertise that would be impossible if everyone needed to know everything.

Research has shown that TMS develops through three interrelated processes: directory updating (learning who knows what), information allocation (assigning responsibilities for acquiring and retaining knowledge), and retrieval coordination (accessing stored information when needed). Lewis (2003) demonstrated that all three processes are essential—teams cannot function effectively with incomplete directories, poorly allocated responsibilities, or inefficient retrieval mechanisms.

For hybrid teams, TMS becomes both more important and more complex. More important because the knowledge landscape expands dramatically: teams now have access not only to human expertise but also to vast AI-encoded knowledge spanning multiple domains, reasoning styles, and levels of granularity. More complex because the 'who knows what' directory must now include entities with radically different knowledge architectures—human experts with deep but narrow domain knowledge, LLMs with broad but shallow cross-domain knowledge, specialized AI tools with focused technical capabilities.

Effective hybrid teams develop what we might call bilateral transactive memory—humans maintain cognitive maps of AI capabilities, and AI agents (to the extent their architectures permit) maintain representations of human expertise patterns. A research team might learn that one particular LLM excels at generating alternative framings, another at grounding claims in empirical literature, and a third at identifying logical inconsistencies. Team members develop shortcuts: when generating hypotheses, consult the creative generalist model; when fact-checking, query the evidence-focused model; when critiquing arguments, engage the analytical model.

This bilateral TMS fundamentally alters team cognition. Instead of a single integrated knowledge base, the team operates as a federated cognitive network—components specialized for different functions, coordinated through meta-knowledge of those specializations. The team's effective intelligence is not the sum of individual intelligences but rather the product of appropriate specialization and efficient knowledge access.

Psychological Safety and Epistemic Safety in Hybrid Teams

Amy Edmondson's (1999, 2003) research on psychological safety has become foundational to understanding team effectiveness. Psychological safety—the shared belief that the team is safe for interpersonal risk-taking—enables teams to surface errors, admit uncertainty, challenge prevailing views, and engage in productive conflict. Without psychological safety, team members self-censor, errors go unreported, learning stagnates, and innovation suffers.

Research across diverse settings—healthcare, aviation, software development, and creative industries—consistently demonstrates that psychological safety predicts team learning, performance, and innovation. The mechanism operates through voice: when team members feel safe, they speak up about problems, ask questions, propose new ideas, and engage in constructive disagreement. This voice behavior, in turn, exposes the team to more information, enables error correction, and facilitates adaptation (Edmondson & Lei, 2014).

For hybrid teams, psychological safety takes on new dimensions. Humans must feel safe to challenge AI outputs, request revisions, reject persuasive but flawed arguments, and admit when they do not understand AI-generated content. This requires what we might call epistemic safety—confidence that questioning machine-generated assertions will not result in negative consequences, and that admitting confusion about AI reasoning is acceptable rather than shameful.

The challenge is significant. Generative AI systems often produce highly confident-sounding outputs even when incorrect or incomplete. They synthesize information fluently, construct seemingly logical arguments, and present conclusions with rhetorical force. This can create subtle pressure on human users to accept outputs uncritically, especially when those users lack domain expertise or feel uncertain about their judgment. The result can be automation bias—the tendency to favor machine-generated suggestions even when those suggestions are erroneous (Goddard et al., 2012).

Establishing epistemic safety in hybrid teams requires several practices. First, teams must normalize critical engagement with AI outputs, treating them as provisional rather than authoritative. Second, team leaders must model questioning behavior, openly identifying AI errors or limitations without deference. Third, teams must develop shared norms around verification—expectations that significant AI-generated claims will be checked against other sources, that unusual results will be examined for plausibility, and that beautiful prose will not substitute for sound reasoning. Fourth, where possible, AI systems themselves should be designed to express appropriate uncertainty, acknowledge limitations, and invite human critique rather than projecting overconfidence.

Interestingly, epistemic safety must also extend to AI systems, in a limited but important sense. Teams benefit when AI agents are designed to dissent from prevailing human views, surface contradictions, ask clarifying questions, and challenge assumptions—provided these behaviors are framed constructively. An AI agent that merely confirms user beliefs or produces outputs designed to maximize user satisfaction will fail as a teammate, because it cannot perform the essential team function of constructive challenge. The team's epistemic environment must permit not only human critique of AI but also AI critique of human reasoning.

Role Differentiation and Adaptive Coordination

A longstanding finding in team research is that effective teams balance role clarity with adaptive flexibility (Marks et al., 2001). Too much role rigidity leads to brittle performance and inability to respond to novel situations; too little role structure produces chaos, duplication, and coordination failure. High-performing teams establish clear initial role definitions while maintaining capacity to adjust those roles dynamically as tasks evolve.

This balance is achieved through what Faraj and Sproull (2000) termed expertise coordination—team members develop shared understanding of each other's knowledge and skills, enabling them to invoke specific expertise when needed and to reconfigure workflows fluidly. In surgical teams, for example, roles are clearly defined (surgeon, anesthesiologist, nurse), yet members constantly adjust their activities based on situational demands, with more senior members sometimes performing tasks typically assigned to juniors and vice versa.

Hybrid teams inherit this need for structured flexibility, but with added complexity. When teams include multiple AI agents alongside human members, role architecture becomes multidimensional. AI agents can be assigned functional roles based on their capabilities: one agent might specialize in hypothesis generation, another in evidence synthesis, a third in critical analysis. Human members can be assigned complementary roles: strategic direction, quality assurance, narrative integration, ethical oversight.

The key insight is that role differentiation in hybrid teams should be grounded in comparative advantage rather than arbitrary assignment. AI agents should perform tasks where their computational power, breadth of knowledge, or freedom from certain cognitive biases provides advantage. Humans should perform tasks requiring judgment, contextual understanding, ethical reasoning, or integration of tacit knowledge. This functional specialization maximizes team

effectiveness while avoiding the pitfall of treating AI as merely faster humans or humans as merely slower AI.

Adaptive coordination becomes even more critical in hybrid configurations. As tasks evolve, teams must dynamically reassign work between human and AI members based on emerging requirements. A research project might begin with extensive AI-driven literature review and hypothesis generation, shift to human-led theoretical refinement and research design, return to AI-assisted data analysis and visualization, and culminate in human-driven interpretation and narrative construction. The team's capacity to orchestrate these handoffs smoothly determines its effectiveness.

Cognitive Diversity and Integration: The Paradox of Variety

One of the most robust findings in team research is that cognitive diversity—variation in knowledge, perspectives, heuristics, and problem-solving approaches—enhances team performance on complex, non-routine tasks (Hong & Page, 2004; Van Knippenberg & Schippers, 2007). Teams whose members think differently explore larger solution spaces, identify more alternatives, catch each other's errors, and avoid groupthink. The benefits of diversity are particularly pronounced for innovation, strategy formulation, and complex problem-solving.

However, diversity is a double-edged sword. While it enhances the raw material available for team cognition, it also complicates integration (Harrison & Klein, 2007). Diverse teams face greater challenges in communication, suffer more frequent misunderstandings, and require more effort to converge on shared interpretations. Without effective integration mechanisms—shared mental models, established communication protocols, strong facilitation—diversity can fragment teams rather than strengthening them.

Hybrid teams naturally possess extreme cognitive diversity. Different LLMs exhibit different 'cognitive styles': some produce expansive, exploratory outputs; others generate concise, structured responses; some emphasize creativity; others prioritize accuracy (Santurkar et al., 2023). When teams engage multiple AI agents, they gain access to this variety of reasoning approaches. Simultaneously, human team members bring their own diversity of expertise, perspective, and judgment. The potential for generative cognitive collision is immense.

Yet this same diversity creates integration challenges that surpass anything traditional teams face. How does a team synthesize highly discursive output from one AI agent with terse, bullet-pointed output from another? How do teams reconcile contradictory framings of the same problem offered by different models? How do they maintain narrative coherence when each AI agent imposes its own organizational logic on content? How do they prevent the proliferation of possibilities from overwhelming decision-making?

The team literature suggests that successful integration requires three conditions. First, teams need integrators—members explicitly tasked with synthesis, translation across perspectives, and convergence toward unified outputs. Second, teams need structured protocols for managing diverse inputs, such as staged evaluation processes or explicit criteria for adjudicating between

alternatives. Third, teams need what might be called interpretive charity—a norm of assuming that apparently contradictory views may be complementary once properly understood, and investing effort in reconciling apparent conflicts before discarding perspectives.

In hybrid teams, these integration mechanisms take on heightened importance. Without skilled human integrators who can synthesize across AI outputs and translate between machine and human cognitive styles, teams risk what we might call cognitive Tower of Babel syndrome—many voices speaking simultaneously in mutually incomprehensible languages, producing noise rather than insight. The integrator role, therefore, becomes one of the most critical functions in hybrid team architecture.

Leadership and the Coherence Anchor

Leadership in teams has traditionally been conceptualized as performing several key functions: setting direction, managing boundaries, obtaining resources, maintaining standards, and providing meaning (Hackman, 2002; Zaccaro et al., 2001). These functions remain relevant in hybrid teams, but a new leadership function emerges with particular force: serving as the coherence anchor.

The concept of coherence anchoring derives from a deeper cognitive role that leadership plays in teams beyond explicit coordination. Leaders, particularly in knowledge work, function as the locus of narrative continuity—the point at which diverse threads of activity, information, and interpretation are woven into coherent storylines. Leaders maintain the 'plot' of the team's work: where the project began, what has been accomplished, what the current obstacles are, where the work is headed, and why it matters. This narrative function is essential for team members to understand their contributions in context and to maintain motivation through inevitable setbacks.

In hybrid teams, the coherence anchor function becomes critical precisely because AI agents have no continuous narrative identity. Each interaction with an AI agent is, from the AI's perspective, essentially independent—the agent does not carry forward a continuous sense of project identity, accumulated wisdom about what has worked, or evolving strategic vision. Only humans maintain true continuity across time. This makes human leaders uniquely positioned to provide the gravitational center around which hybrid team cognition organizes itself.

Consider what happens in a hybrid team lacking strong coherence anchoring. Multiple AI agents generate outputs responding to different prompts; each output is locally coherent but globally disconnected from the others. Different team members pursue different threads, guided by different AI-generated suggestions. The team produces volume but not synthesis. Ideas proliferate but do not accumulate. The work exhibits what might be called high entropy—lots of activity and information, but little structure or direction.

Effective hybrid team leaders counter this tendency toward entropy through active coherence maintenance. They articulate the overarching vision repeatedly. They explicitly connect AI-generated outputs to ongoing project threads. They identify which of many possible directions the team will pursue and which will be deferred. They ensure that insights from one phase of work inform the next phase. They serve, in effect, as the team's working memory—maintaining

continuity that enables the team to function as a learning system rather than a collection of disconnected episodes.

This coherence anchor function suggests that leadership in hybrid teams requires somewhat different capabilities than leadership in traditional teams. Beyond standard leadership competencies—strategic thinking, interpersonal skill, decision-making—hybrid team leaders need what might be called narrative intelligence: the ability to maintain thematic consistency across varied and voluminous inputs, to recognize patterns connecting disparate fragments, to distinguish signal from noise in high-information environments, and to construct compelling storylines that give meaning to distributed cognitive processes. These are capabilities that, at present, only humans can provide.

III. POSITIONING WITHIN EXISTING RESEARCH: CONTRIBUTIONS AND DIFFERENTIATIONS

Having established the organizational behavior foundations, we now position our framework within the broader research landscape on human–AI systems. This positioning is essential for specifying our theoretical contribution: what exists in prior literature, where gaps remain, and how our framework extends and synthesizes existing knowledge.

The Hybrid Intelligence Research Stream

The term "hybrid intelligence" has been employed in multiple research contexts with varying meanings. Dellermann et al. (2019) define hybrid intelligence as the combination of human and artificial intelligence to achieve superior outcomes to what either could accomplish alone, with particular emphasis on complementary task allocation. This perspective, emerging from information systems research, focuses primarily on dyadic human–AI configurations and emphasizes design principles for effective collaboration.

Rahwan et al. (2019) introduced the concept of "machine behaviour" as a framework for studying AI systems as behavioral subjects in their own right, analogous to how behavioral sciences study animals or humans. This work provides a powerful framing for treating AI agents as entities with observable behavioral patterns, preferences, and response tendencies—a perspective that aligns with our treatment of AI agents as team members with distinct cognitive profiles.

Kamar (2016) articulated foundational principles for hybrid human–machine computation, emphasizing complementary capabilities, appropriate task allocation, and the importance of transparency in human–AI systems. This work established key principles that inform our framework, particularly regarding the necessity of mutual understanding between human and artificial team members.

Our framework extends this hybrid intelligence literature in three ways. First, we move from dyadic (one human, one AI) to team configurations (multiple humans, multiple AI agents). Second, we integrate organizational behavior theory systematically rather than treating human coordination as secondary to technical design. Third, we specify mechanisms—shared mental models, transactive memory, epistemic safety—that enable hybrid team effectiveness, providing testable constructs rather than general principles.

Distributed Cognition and Extended Mind Traditions

The distributed cognition framework (Hutchins, 1995; Hollan et al., 2000) has long argued that cognitive processes are not confined to individual minds but are distributed across people, artifacts, and environmental structures. Hutchins' classic study of navigation teams demonstrated how knowledge and reasoning are socially distributed, with the team functioning as a cognitive system whose properties exceed those of individual members.

The extended mind thesis (Clark & Chalmers, 1998) makes a stronger claim: cognitive processes can extend beyond the brain to include external tools and artifacts that function as integral parts of the cognitive system. When an individual uses a notepad to perform calculations, the notepad becomes part of the cognitive apparatus, not merely a tool used by an independent mind.

Our framework builds on these traditions by treating hybrid intelligence teams as genuinely distributed cognitive systems. However, we extend this work in important ways. First, prior distributed cognition research rarely addressed systems where some cognitive components are autonomous AI agents with their own reasoning capabilities. Second, we specify coordination mechanisms (bilateral TMS, cross-species SMMs) required when cognitive distribution spans human and artificial substrates with fundamentally different architectures. Third, we identify failure modes (epistemic drift, coherence collapse) specific to hybrid cognitive systems.

Human–AI Teaming and Human–Autonomy Interaction

Research on human–robot teams and human–autonomy interaction has explored how humans coordinate with semi-autonomous systems in high-stakes domains: military operations, aviation, manufacturing, and emergency response (Chen et al., 2014; Endsley, 2017). This literature emphasizes trust calibration, transparency, situation awareness, and appropriate reliance as key factors in effective human–autonomy teaming.

Johnson et al. (2016) proposed design principles for human–machine teamwork, emphasizing the importance of common ground, predictability, and observability. Demir et al. (2019) demonstrated how team situation awareness frameworks from organizational behavior can be extended to human–autonomy contexts, showing that shared awareness of goals, roles, and environmental conditions predicts team effectiveness.

Our framework builds on these insights while addressing important differences between hybrid intelligence teams and prior human–autonomy teaming research. First, generative AI systems are fundamentally different from the narrow automation studied in most human–autonomy research—LLMs engage in open-ended reasoning rather than executing predefined procedures.

Second, we address multi-agent configurations where coordination occurs not only between humans and AI but also across multiple AI agents with different capabilities. Third, we focus on knowledge work contexts where the primary challenge is integration and synthesis rather than real-time coordination under time pressure.

Human–Computer Interaction and Mixed-Initiative Systems

Human–computer interaction research, particularly work on mixed-initiative systems (Horvitz, 1999), addresses scenarios where both human and machine can take initiative in problem-solving. This bi-directional influence captures aspects of generative AI collaboration more effectively than command-and-control models.

Shneiderman (2020) articulated principles for human-centered AI, emphasizing the importance of maintaining human control, ensuring algorithmic accountability, and designing for augmentation rather than automation. Amershi et al. (2019) provided guidelines for human–AI interaction that emphasize clarity about system capabilities and limitations, support for efficient invocation and dismissal, and appropriate time-scoping of AI suggestions.

While this HCI work provides valuable design principles, it predominantly addresses dyadic interaction (one user, one system) and typically treats AI as a tool invoked by humans rather than as a team member contributing to ongoing collaboration. Our framework extends HCI insights by specifying how multiple humans and multiple AI agents coordinate as integrated teams, how role architectures emerge in multi-agent configurations, and how team-level properties (shared mental models, transactive memory) enable sustained collaboration beyond individual human–AI interactions.

Multi-Agent AI Systems Research

Research on multi-agent AI systems (Wooldridge, 2009; Stone & Veloso, 2000) addresses coordination among multiple AI agents. This work has produced sophisticated frameworks for agent communication protocols, task decomposition, negotiation, and distributed problem-solving.

Recent work has explored multi-LLM configurations, including debate protocols where multiple models argue different positions (Du et al., 2023), ensemble methods combining outputs from diverse models, and specialized agent architectures for complex tasks. This research demonstrates that interactions between multiple AI agents can produce emergent behaviors and improved outcomes beyond single-model performance.

However, multi-agent AI research typically assumes all agents are computational—there are no human team members. This eliminates several complexities central to hybrid teams: the need for narrative coherence and meaning-making, psychological and epistemic safety, cross-species communication challenges, and ethical dimensions of decision-making. Our framework synthesizes insights from multi-agent AI with organizational behavior theory to address the sociotechnical system that emerges when humans and multiple AI agents collaborate.

AI Alignment and Value Alignment Literature

Research on AI alignment (Amodei et al., 2016; Gabriel, 2020) addresses the challenge of ensuring AI systems pursue objectives aligned with human values and intentions. This work emphasizes oversight mechanisms, error detection, reward specification, and interpretability.

Alignment concerns become particularly salient in hybrid teams. When multiple AI agents contribute to decision-making, ensuring that their combined influence aligns with team values and organizational goals becomes more complex. Our constructs of epistemic safety and judgment authority address alignment challenges at the team level: epistemic safety ensures humans feel empowered to challenge AI outputs, while judgment authority specifies that consequential decisions requiring ethical reasoning remain under human control.

The Critical Gap: Integrative Theory for Multi-Human, Multi-AI Teams

Across these research streams, a consistent pattern emerges. Each literature illuminates specific aspects of human–AI systems: hybrid intelligence emphasizes complementarity, distributed cognition emphasizes social distribution of reasoning, human–autonomy teaming emphasizes trust and transparency, HCI emphasizes interaction design, multi-agent AI emphasizes coordination protocols, and alignment research emphasizes value alignment. Yet none provides an integrative framework for the configuration that is becoming increasingly common in practice: teams consisting of multiple humans and multiple AI agents, engaged in extended collaboration on complex, open-ended problems.

The gap is not merely empirical—a lack of studies on the right configurations. It is theoretical. Existing frameworks do not specify how hybrid teams should be structured, what mechanisms enable their effectiveness, what failure modes they face, and how they differ fundamentally from both all-human teams and all-AI multi-agent systems. We lack models that integrate insights from organizational behavior with insights from AI research in ways that generate testable propositions and guide practice.

This theoretical vacuum has practical consequences. Organizations deploying AI in team contexts operate without clear guidance on team design. They lack frameworks for deciding how many AI agents a team should incorporate, what roles those agents should play, how to structure human–AI interaction, or what leadership practices enable coordination. They cannot anticipate the failure modes specific to hybrid teams or develop countermeasures in advance.

What is needed, therefore, is integrative theory-building—the development of frameworks that can organize existing insights, generate testable predictions, and guide practice. The next section undertakes this work, proposing a comprehensive framework for understanding hybrid intelligence teams as a distinctive category of cognitive system.

IV. THEORY BUILDING: THE HYBRID INTELLIGENCE TEAM FRAMEWORK

Definitions and Scope

A **hybrid intelligence team (HIT)** is defined as a work group composed of two or more humans and two or more AI agents, engaged in interdependent cognitive tasks over time, with the explicit goal of producing integrated outputs that require contributions from both human and artificial members.

This definition establishes several boundary conditions:

- **Genuinely collaborative**: HITs require integration of human and AI contributions rather than mere parallel processing
- **Multiple AI agents**: HITs involve multiple AI agents with potentially different capabilities, not just multiple instances of the same AI performing repetitive tasks
- **Persistent over time**: HITs allow for learning, adaptation, and development of team-specific coordination patterns
- **Synthetic outputs**: HITs produce co-created outputs where the division between human and AI contribution becomes meaningless

What HITs are not:

- A single analyst working with one AI assistant (dyadic, not team configuration)
- Automated workflows with only occasional human oversight (insufficient human-AI interdependence)
- Short-lived micro-collaborations lasting minutes (insufficient persistence for team development)
- Simple aggregation of independent human and AI outputs without integration

Scope of AI "agents": For our purposes, an AI agent is a computational system capable of autonomous generation of content, reasoning, or analysis in response to human direction. This includes distinct foundation models with different capabilities (e.g., GPT-4, Claude, Gemini), specialized fine-tuned versions of foundation models, and retrieval-augmented generation systems. Multiple instances of the same model performing identical tasks would not constitute multiple agents in our framework.

The Extended IMOI Model for Hybrid Teams

Our framework builds on the Input-Mediators-Outputs-Inputs (IMOI) model widely used in team research (Ilgen et al., 2005). The traditional IMOI model specifies that team inputs (composition, resources, design) shape mediating processes (coordination, communication, conflict), which produce outputs (performance, innovation, member satisfaction), which then feed back as inputs for subsequent team cycles.

For hybrid teams, we extend the IMOI model to accommodate the distinctive properties of human–AI collaboration:

Table 1: Extended IMOI Model for Hybrid Intelligence Teams

Component	Traditional Human Teams	Hybrid Intelligence Teams	Key Extensions
Inputs	Human member characteristics (expertise, personality, demographics); Team design (size, structure); Task characteristics; Organizational context	All traditional inputs PLUS: AI agent profiles (reasoning style, knowledge base, interaction patterns); AI composition (number, diversity of agents); Human metacognitive capability; AI literacy	Addition of AI-specific inputs; Recognition that both human AND AI characteristics shape potential
Mediators	Shared mental models; Transactive memory; Communication patterns; Psychological safety; Conflict management; Leadership behaviors	All traditional mediators PLUS: Cross-species shared mental models; Bilateral transactive memory; Epistemic safety; Coherence anchoring; Human-AI integration protocols; Role fluidity mechanisms	New hybrid-specific constructs that bridge human and AI cognition; Explicit coherence maintenance
Outputs	Task performance; Innovation; Efficiency; Team viability; Member satisfaction	All traditional outputs PLUS: Co-created knowledge artifacts; Evolved coordination routines; Meta-learning about optimal collaboration patterns; Dual learning trajectories (human and AI adaptation)	Recognition of learning at multiple levels; Artifacts as distinct output category
Feedback Loops	Team learning from experience; Adaptation of routines; Member socialization	All traditional loops PLUS: AI fine-tuning or prompt refinement based on team experience; Bilateral adaptation (humans learning about AI AND AI adapting to humans); Potential for steeper learning curves	More direct and rapid feedback loops possible; Bilateral adaptation mechanisms

This extended model generates several testable hypotheses:

- **H1**: Teams with higher levels of cross-species SMM accuracy will demonstrate superior coordination efficiency compared to teams with lower SMM accuracy, controlling for technical AI capabilities.

- **H2**: The relationship between AI cognitive diversity and team performance will be curvilinear, with moderate diversity optimal and very high diversity requiring disproportionate integration effort.
- **H3**: Epistemic safety will mediate the relationship between team psychological safety and the team's ability to identify and correct AI-generated errors.
- **H4**: The quality of coherence anchoring will predict team-level narrative consistency over time, independent of the sophistication of AI agents employed.

Core Constructs: Formal Definitions and Measurement

We now provide formal definitions for the key constructs in our framework, along with observable indicators and proposed measurement approaches.

Cross-Species Shared Mental Models

Definition: Cross-species shared mental models (CS-SMM) are cognitive representations of task structure, role definitions, quality standards, and interaction protocols that exhibit overlap between human team members and AI agents, despite fundamental differences in cognitive architecture. CS-SMMs enable implicit coordination by providing common reference points for interpreting situations and anticipating actions.

Observable indicators:

- Consistency in how humans and AI agents decompose complex tasks
- Alignment in quality standards applied to outputs
- Coordination efficiency measured by reduced need for explicit communication
- Accuracy of human predictions of AI agent responses
- Appropriateness of AI responses to implicit human cues

Measurement approaches:

1. *Similarity assessment*: Human team members and AI agents independently describe task structure, role expectations, and quality criteria; overlap is quantified through semantic similarity metrics (cosine similarity of embeddings) or structured coding schemes
2. *Coordination efficiency metrics*: Time to complete coordinated activities, frequency of coordination failures, number of clarification requests
3. *Predictive accuracy*: Humans predict AI agent outputs before generation; accuracy of predictions indicates SMM quality
4. *Expert evaluation*: Domain experts rate the degree of shared understanding evident in team interactions

Hypothesized relationships:

- CS-SMM accuracy should correlate positively with team coordination efficiency
- CS-SMM develops over time through repeated interaction

- Explicit SMM development interventions (structured discussions of task models) should accelerate CS-SMM formation

Bilateral Transactive Memory

Definition: Bilateral transactive memory (BTM) is a meta-knowledge structure in which (a) human team members maintain cognitive maps of AI agent specializations, capabilities, and limitations, and (b) AI agents (to the extent their architectures permit) maintain representations of human expertise, preferences, and priorities. BTM enables efficient distributed knowledge access across human and AI team members.

Observable indicators:

- Accuracy of knowledge access: Do team members consistently direct queries to the appropriate human or AI source?
- Response time to knowledge requests
- Redundancy in knowledge acquisition (lower redundancy indicates better BTM)
- Utilization patterns: Are specialized AI capabilities fully leveraged or underutilized?
- Adaptation to personnel/agent changes

Measurement approaches:

1. *Directory accuracy assessment*: Test team members' knowledge of "who/what knows what" through structured questionnaires; compare to objective measures of actual expertise
2. *Network analysis*: Map information flow within teams; efficient BTM should show appropriate specialization and minimal redundant knowledge access
3. *Efficiency metrics*: Time from query initiation to authoritative answer; path length in knowledge access networks
4. *Adapted TMS scales*: Modify Lewis (2003) TMS scales to include AI agents; assess specialization, credibility, and coordination dimensions

Hypothesized relationships:

- BTM quality should predict efficiency of knowledge-intensive tasks
- Teams with explicit AI capability documentation should develop BTM faster
- BTM should moderate the relationship between AI diversity and team performance (high BTM enables leverage of diversity)

Epistemic Safety

Definition: Epistemic safety is the shared team belief that questioning, challenging, and revising AI-generated content is not only permitted but expected and valued. Epistemic safety represents an extension of psychological safety specific to human–AI contexts, focused on the willingness to exercise critical judgment over machine-generated outputs despite their apparent authority or fluency.

Observable indicators:

- Frequency of human challenges to AI outputs
- Willingness to request AI output revision or regeneration
- Expression of uncertainty about AI-generated content
- Instances of overriding AI recommendations
- Quality of critical evaluation (substantive critique vs. uncritical acceptance)
- Absence of automation bias indicators

Measurement approaches:

1. *Behavioral coding*: Code team interactions for instances of AI output challenge, modification, or rejection; calculate rate relative to total AI outputs
2. *Survey instruments*: Adapt psychological safety scales (Edmondson, 1999) with items specific to AI interaction (e.g., "I feel comfortable questioning AI outputs even when they seem authoritative"; "Team members do not judge each other for expressing confusion about AI reasoning")
3. *Critical thinking quality assessment*: Expert evaluation of the depth and appropriateness of human critique of AI outputs
4. *Error detection rates*: Introduce known errors in AI outputs; measure detection and correction rates as indicators of epistemic safety in practice

Hypothesized relationships:

- Epistemic safety should predict error detection and correction rates
- Leader modeling of AI critique should strengthen team epistemic safety
- Epistemic safety should mediate the relationship between AI output fluency and critical evaluation quality

Coherence Anchoring

Definition: Coherence anchoring is the process through which designated human team members (typically leaders) maintain narrative continuity, thematic consistency, and strategic direction across diverse AI-generated outputs and team activities. The coherence anchor serves as the integrative center that prevents high-entropy fragmentation of team cognition.

Observable indicators:

- Thematic consistency across team outputs over time
- Frequency of explicit coherence maintenance activities (vision articulation, integration discussions)
- Stability of project direction versus drift
- Quality of narrative integration in final outputs
- Team member alignment on project goals and status

Measurement approaches:

1. *Computational text analysis*: Assess semantic coherence across team documents over time using topic modeling, semantic similarity metrics, and drift detection algorithms
2. *Process observation*: Code team interactions for coherence maintenance behaviors (e.g., "As we discussed last week...", "This connects to our earlier finding that...", "This doesn't fit with our main narrative")
3. *Expert evaluation*: Independent judges rate narrative quality and thematic consistency of team outputs
4. *Team member surveys*: Assess perceived clarity of direction, understanding of how activities connect to goals
5. *Longitudinal goal stability*: Track consistency of team objectives and framings over time

Hypothesized relationships:

- Quality of coherence anchoring should predict narrative consistency of outputs
- Teams with designated coherence anchors should outperform teams lacking this role
- Coherence anchoring should moderate the relationship between AI output volume and integration quality

Additional Constructs Requiring Formal Development

Beyond the four constructs defined above, our framework introduces several additional concepts that require formal specification in future work:

- **Cognitive abundance overload**: Information overload arising not from low-quality noise but from excess high-quality alternatives
- **Epistemic drift**: Gradual unintended shift in team understanding, framing, or standards driven by accumulated AI influence
- **Narrative fragmentation**: Inability to integrate multiple AI outputs with different organizational logics into coherent wholes
- **False convergence**: Mistaking consistency across AI outputs (potentially due to correlated training) for truth
- **Coherence collapse**: Catastrophic failure where team loses all thematic unity despite continued activity
- **Authority confusion**: Inability to clearly delineate human versus AI responsibility for decisions

Each of these constructs represents a potential failure mode and should be operationalized for empirical investigation.

Role Architecture: Functional Differentiation in Hybrid Teams

Effective HITs require thoughtful role architecture—the distribution of functions across human and AI team members based on comparative advantage. We propose role structures for both AI agents and human members.

AI Agent Roles

Generator agents specialize in divergent thinking: producing alternatives, exploring possibility spaces, and generating novel framings. These agents are optimized for breadth over precision, creativity over accuracy. They expand the team's option set and prevent premature convergence. Implementation often involves models with higher temperature settings or explicitly prompted for creative exploration.

Critic agents specialize in analytical evaluation: identifying logical flaws, surfacing unstated assumptions, and challenging conclusions. These agents provide constructive skepticism essential for quality control, playing the role of devil's advocate without the social friction that can inhibit human critique. Implementation may involve prompts that explicitly request identification of weaknesses, alternative perspectives, or logical inconsistencies.

Synthesizer agents specialize in integration: combining disparate pieces into coherent wholes, identifying connections across domains, and producing structured summaries. These agents help manage information overload by distilling voluminous inputs into digestible forms. Implementation typically involves prompts requesting integration of multiple sources or creation of structured frameworks organizing diverse material.

Verifier agents specialize in fact-checking and evidence grounding: validating claims against established knowledge, identifying contradictions, and flagging unsupported assertions. These agents provide a bulwark against hallucination and ensure that outputs remain empirically grounded. Implementation may involve retrieval-augmented generation architectures or explicit verification protocols.

Coherence-checker agents specialize in consistency maintenance: identifying drift from established frameworks, flagging contradictions across outputs, and ensuring logical alignment. These agents support the human coherence anchor by performing computational consistency checks that would be cognitively taxing for humans. Implementation involves prompts that compare current outputs to earlier content and identify discrepancies.

Note on implementation: These roles may be instantiated as different foundation models with distinct capabilities, as different configurations of the same model via prompt engineering, or as a combination. The key is functional differentiation rather than technical separation.

Human Roles in Hybrid Teams

Human team members fulfill complementary functions that AI cannot adequately perform:

The coherence anchor (typically the team leader or principal investigator) maintains narrative continuity, strategic direction, and thematic consistency. This role involves repeatedly articulating the project's core purpose, connecting current activities to broader goals, and ensuring that diverse AI outputs serve unified ends. Required capabilities include narrative intelligence, strategic vision, and synthesis skill.

The judgment authority makes consequential decisions that require ethical reasoning, stakeholder consideration, or tacit domain knowledge. While AI can inform these decisions,

ultimate authority remains human—a crucial feature for maintaining accountability and ensuring that decisions reflect organizational values and contextual nuance. Required capabilities include domain expertise, ethical reasoning, and decisiveness.

The integration specialist synthesizes across AI outputs, translates between human and AI cognitive styles, and resolves apparent contradictions by identifying deeper complementarities. This role requires high tolerance for ambiguity, skill at pattern recognition across diverse representations, and capacity to construct coherent narratives from fragmented inputs. Required capabilities include cognitive flexibility, synthesis ability, and metacognitive skill.

The quality steward maintains standards, evaluates AI outputs for accuracy and appropriateness, and decides when outputs are sufficient versus when further iteration is required. This role demands domain expertise, critical thinking skill, and resistance to automation bias—the ability to reject plausible-sounding but ultimately inadequate AI-generated content. Required capabilities include domain expertise, critical evaluation skill, and epistemic confidence.

Failure Modes and Countermeasures

Understanding failure modes is essential for theory development, as it specifies boundary conditions and points toward necessary countermeasures. HITs face several distinctive pathologies beyond those affecting traditional teams.

Cognitive Overload Through Abundance

Description: Cognitive overload through abundance occurs when AI systems generate options, analyses, or information faster than humans can meaningfully process. Unlike traditional information overload (too much low-quality data to examine), abundance overload involves too many plausible alternatives to evaluate adequately. Teams drown not in noise but in signal—everything seems potentially valuable, making prioritization paralyzing.

Observable indicators: Increasing cycle time for decisions as options multiply; paralysis in choice contexts; expressions of feeling overwhelmed despite high-quality inputs; declining decision quality as option set expands.

Countermeasures:

- Establish explicit constraints on AI generation (e.g., "provide three strongest options" rather than exhaustive alternatives)
- Implement staged evaluation: AI generates broadly, humans cull to top candidates, AI elaborates finalists
- Develop clear prioritization criteria in advance of generation

Theoretical connection: This failure mode connects to bounded rationality (Simon, 1971) and attention as a scarce resource. Empirical investigation should examine the relationship between option set size and decision quality/speed.

Epistemic Drift

Description: Epistemic drift occurs when a team gradually shifts its understanding, framing, or standards without deliberate intention, pulled by the accumulated weight of AI-generated content. Each AI output subtly influences subsequent prompts; each synthesis alters the context for the next interaction. Over time, the team may pursue objectives or employ frameworks quite different from initial intentions, not through conscious choice but through drift.

Observable indicators: Divergence between initial and current project framing; inability to reconstruct reasoning for key decisions; disconnection from original goals; changing standards for evidence or quality without explicit discussion.

Countermeasures:

- Maintain explicit version control and lineage tracking for key decisions and framings
- Conduct periodic "coherence audits" comparing current direction to original intentions
- Document reasoning behind significant pivots or frame changes
- Schedule regular human-only integration sessions to assess trajectory

Theoretical connection: This phenomenon parallels organizational drift (Vaughan, 1996) but operates at faster timescales due to the pace of AI-mediated interaction.

Narrative Fragmentation

Description: Narrative fragmentation happens when multiple AI agents impose different organizational logics on content, and these logics prove difficult to reconcile. One agent structures content chronologically, another thematically, a third by logical dependency. The resulting outputs, while individually coherent, resist integration into unified narratives. Teams spend excessive effort on reconciliation rather than progress.

Observable indicators: Inconsistent organizational structures across documents; difficulty integrating outputs; time spent on reconciliation exceeding generation; team member confusion about overall structure.

Countermeasures:

- Establish consistent structural templates before AI generation
- Assign a single agent (or human) as primary structurer
- Conduct integration at deliberate milestones rather than attempting continuous synthesis
- Provide all agents with explicit structural requirements

Theoretical connection: This failure mode relates to cognitive load theory (Sweller, 1988) and the challenges of integrating multiple representational formats.

False Convergence

Description: False convergence occurs when teams mistake consistency across AI outputs for truth. If multiple AI agents produce similar responses (perhaps because they are trained on similar data or employ similar reasoning heuristics), teams may over-weight that consensus. Unlike human groupthink, where social dynamics drive premature agreement, false convergence arises from statistical correlation in AI training—a subtler and potentially more insidious phenomenon.

Observable indicators: High confidence based on AI output agreement; insufficient external verification when AI outputs align; surprise when external evidence contradicts AI consensus.

Countermeasures:

- Recognize that AI consensus is not independent evidence
- Maintain skepticism toward convergent AI outputs, particularly on contentious questions
- Explicitly seek external verification when multiple AI agents agree
- Employ AI agents with deliberately different training data or architectures

Theoretical connection: This parallels groupthink (Janis, 1972) but operates through different mechanisms. Research should examine whether standard groupthink countermeasures (e.g., devil's advocate roles) translate to AI contexts.

Coherence Collapse

Description: Coherence collapse represents a catastrophic failure mode where the team loses all thematic unity. Without effective coherence anchoring, work devolves into disconnected episodes: each interaction with AI produces locally reasonable outputs that do not accumulate into larger understanding. The team exhibits activity without progress, generating volume without synthesis.

Observable indicators: Inability to summarize project status or findings; disconnected outputs that do not build on each other; circular conversations revisiting settled questions; lack of cumulative progress despite extensive activity.

Countermeasures:

- Designate explicit coherence anchor role
- Conduct regular synthesis sessions led by coherence anchor
- Maintain living documents that capture cumulative understanding
- Establish project milestones requiring integrated outputs

Theoretical connection: This failure mode represents an entropy concept from information theory applied to team cognition—without active energy input (coherence maintenance), systems tend toward disorder.

Authority Confusion

Description: Authority confusion occurs when teams cannot clearly delineate human versus AI responsibility for decisions. If AI systems provide recommendations with high confidence, humans may defer inappropriately on judgments that should remain human. Conversely, humans may overrule AI on technical matters where AI possesses superior knowledge.

Observable indicators: Ambiguity about who/what made key decisions; inability to assign accountability for outcomes; inappropriate deference to or dismissal of AI recommendations; post-hoc rationalization of decisions.

Countermeasures:

- Establish explicit authority structures specifying decision types reserved for humans
- Document decision provenance (AI recommendation vs. human judgment)
- Create clear protocols for human override of AI suggestions
- Train team members in appropriate reliance (when to trust vs. verify AI)

Theoretical connection: This failure mode connects to accountability research in organizations and to algorithmic accountability in AI ethics literature (Gabriel, 2020).

Mechanisms for Maintaining Effectiveness

To counter the failure modes identified above and enable effective functioning, HITs must implement specific practices and structural features:

Staged Integration Protocols

Teams establish deliberate convergence points where diverse AI outputs are synthesized before proceeding. Rather than continuously accumulating AI-generated content, teams create rhythm: divergent exploration phases followed by integrative consolidation phases. This rhythm prevents unbounded accumulation while maintaining generative capacity.

Explicit Meta-Discussions

Periodically, teams step back from content production to discuss their collaboration process itself: Are we integrating AI outputs effectively? Is our coherence anchor functioning adequately? Are we challenging AI outputs appropriately? Such meta-cognitive reflection enables teams to detect and correct dysfunction before it becomes entrenched.

Version Control and Lineage Tracking

Because content evolves through multiple AI-mediated iterations, teams must maintain explicit records: where ideas originated, why certain framings were adopted, what alternatives were considered. This documentation helps prevent epistemic drift and enables teams to backtrack when they detect errors.

Prompting Protocols

Effective teams develop shared practices for AI interaction: always request critique alongside generation, ask AI to identify limitations or uncertainties, request alternative framings before converging, explicitly probe for contradictions with existing content. These protocols transform ad hoc AI use into disciplined team practice.

Human-Only Synthesis Sessions

Periodically, human team members convene without AI to consolidate understanding, make strategic decisions, and perform narrative integration requiring tacit knowledge or ethical judgment. This creates a dual-layer team structure: a full hybrid team for production, and a human core team for high-level synthesis and direction-setting.

Quality Thresholds and Acceptance Criteria

By establishing explicit standards in advance—what constitutes adequate evidence, what level of logical rigor is required, what stylistic consistency must be maintained—teams create benchmarks against which to evaluate AI-generated content. This converts subjective judgment into more objective assessment.

V. THE RESEARCH AGENDA: ESTABLISHING HYBRID INTELLIGENCE TEAMS AS A SCHOLARLY FIELD

The framework proposed in Section IV generates numerous testable propositions, points toward critical measurement challenges, and suggests new methodological approaches. This section articulates a comprehensive research agenda for the emerging field of hybrid intelligence teams.

Core Theoretical Questions

Several fundamental theoretical questions require systematic investigation:

Performance boundaries and moderators: Under what conditions do HITs outperform all-human teams, and what factors moderate this relationship? We hypothesize that HITs excel when tasks involve high information processing demands, benefit from cognitive diversity, and require both creative generation and rigorous evaluation—but empirical verification across domains is necessary. Conversely, some tasks may suffer from the coordination overhead hybrid configurations impose.

Team size and AI composition: How does team size interact with AI composition? Traditional team research identifies optimal sizes (generally 4-7 members for knowledge work), but how does this change when some members are AI? Does including three AI agents alongside two humans create a psychologically five-person team, or does the human/AI ratio matter more than raw team size?

Optimal AI cognitive diversity: What determines the optimal level of AI cognitive diversity in teams? While diversity generally benefits human teams up to a point (Van Knippenberg & Schippers, 2007), does the same principle apply to AI agents? Should teams employ multiple instances of the same model (for consistency) or different models with varied cognitive profiles (for diversity)?

Shared mental model development: How do CS-SMMs develop in HITs, and how complete must they be for effective coordination? Given that perfect mutual understanding between humans and AI is likely impossible, what minimum level of shared representation suffices?

Leadership behaviors: What leadership behaviors most effectively maintain coherence in HITs? Do traditional leadership practices translate to hybrid contexts, or do HITs require distinctive practices—for instance, continuous narrative reinforcement, active integration of AI outputs, or explicit coherence monitoring?

Epistemic safety emergence: How does epistemic safety emerge and stabilize in HITs? Does questioning AI require the same mechanisms as psychological safety, or do new practices prove more effective?

Measurement Priorities

Studying HITs requires new measurement approaches and instruments. Key priorities include:

Cross-species shared mental model assessment: Develop validated measures combining similarity assessment (semantic analysis of task descriptions), coordination efficiency metrics, and predictive accuracy approaches.

Bilateral transactive memory measurement: Adapt Lewis (2003) TMS scales to include AI agents; validate through network analysis of information flow and efficiency metrics.

Epistemic safety scales: Create validated survey instruments specific to human–AI interaction; complement with behavioral coding of AI output challenge frequency and quality.

Coherence quality metrics: Develop computational approaches (semantic coherence analysis, drift detection algorithms) validated against expert evaluation of narrative quality.

Failure mode detection: Create validated instruments for detecting cognitive abundance overload, epistemic drift, narrative fragmentation, false convergence, coherence collapse, and authority confusion.

Longitudinal team development measures: Track evolution of coordination routines, shared mental models, and performance over time; current team research is predominantly cross-sectional.

Methodological Innovations

Studying HITs demands methodological innovation beyond standard team research approaches:

Digital ethnography: Because human–AI interaction occurs through digital interfaces, researchers can capture complete interaction logs: every prompt, every AI response, every revision. This comprehensive data enables fine-grained analysis of coordination patterns, communication dynamics, and workflow evolution. Combining logs with periodic interviews triangulates objective behavioral patterns with subjective experience.

Controlled experiments with systematic AI variation: Manipulate AI agent characteristics—number, cognitive diversity, role specifications—while holding task and team composition constant. Random assignment to hybrid configurations enables causal inference. Such experiments might be conducted as laboratory studies or field experiments within organizations.

Simulation studies using LLMs: Create synthetic hybrid teams—human participants collaborating with multiple AI agents—and systematically vary AI behavior to test theoretical predictions. While external validity questions remain, simulations enable testing of theoretically important conditions difficult to achieve in natural settings.

Longitudinal case studies: Deep qualitative investigation of real hybrid teams engaged in extended projects can reveal team evolution: how coordination routines develop, how teams respond to failures, what informal practices emerge, how shared mental models crystallize over repeated interaction.

Computational analysis of interaction patterns: Natural language processing can reveal coordination signals: who defers to whom, who questions whose outputs, linguistic markers of uncertainty or confidence. Social network analysis can map information flow, revealing whether teams exhibit efficient transactive memory. Machine learning models trained on successful versus unsuccessful hybrid teams might identify predictive features not obvious to human observers.

Comparative cross-domain studies: Examine HITs across different domains (software development, scientific research, strategic consulting, creative production) to identify universal versus domain-specific principles.

Theoretical Extensions and Future Directions

Several research directions appear particularly promising for advancing hybrid team theory:

Team learning and adaptation theory: How do teams develop increasingly efficient coordination routines? How does team-level learning differ from individual learning about AI? What feedback mechanisms enable rapid adaptation versus entrench dysfunctional patterns? Current team learning theory (Edmondson, 1999) may require extension to accommodate bilateral learning (humans about AI, AI adaptation to humans).

Multi-level theory: Connect individual, team, and organizational phenomena. How do individual AI literacy and metacognitive skill aggregate to produce team-level capabilities? How

do team-level coordination routines diffuse across organizational units? How do organizational structures and cultures enable or constrain effective hybrid collaboration? Multi-level models can specify micro-foundations of hybrid team effectiveness while situating teams in broader organizational contexts.

Cross-cultural comparative research: Examine how hybrid teams function across cultures and institutional contexts. Does individualism-collectivism shape comfort with AI teammates? Do power distance norms affect willingness to challenge AI outputs? Do different professional cultures develop distinctive approaches to hybrid collaboration?

Contingency theory: Specify when hybrid teams excel versus when traditional teams suffice. What task characteristics, environmental conditions, and organizational capabilities moderate hybrid team effectiveness? This can guide organizations in making appropriate team design choices.

Co-evolutionary theory: Address how hybrid teams and AI systems co-evolve. As teams develop sophisticated collaboration practices, they demand more capable AI teammates. As AI systems become more sophisticated, they enable new forms of collaboration. This co-evolutionary dynamic may produce step-changes in knowledge work organization.

VI. IMPLICATIONS FOR PRACTICE, POLICY, AND SOCIETY

Beyond advancing theory, research on HITs has profound implications for organizational practice, educational systems, and broader societal structures.

Organizational Design and Management

Team design capabilities: Human resource professionals must learn to assess not only human member fit but also optimal AI composition. This involves understanding AI agent capabilities, cognitive profiles, and interaction patterns—knowledge currently concentrated among technical specialists. Organizations should develop hybrid team design guidelines: decision frameworks for determining how many and which AI agents to include, role architecture templates adapted to different task types, and protocols for evaluating hybrid team effectiveness.

Leadership development: Leadership training must evolve to emphasize hybrid-specific competencies. Traditional programs focus on interpersonal skills, strategic thinking, and decision-making. Hybrid team leaders additionally require metacognitive skill (evaluating AI outputs critically), integration capability (synthesizing across cognitive styles), narrative intelligence (maintaining coherence), and cognitive flexibility (comfort operating where definitive answers are scarce). Leadership development programs should incorporate these dimensions explicitly.

Performance management: How should organizations attribute performance to human versus AI contributions? Should teams be evaluated on raw output quality, efficiency of human–AI coordination, or learning trajectory? Should individuals be assessed on their ability to leverage AI effectively, their contribution to team integration, or both? Organizations must develop frameworks that fairly recognize human contributions while acknowledging AI's role.

Knowledge management adaptation: Traditional knowledge management focuses on codifying human expertise. In hybrid contexts, organizations should also document effective AI prompting strategies, successful integration protocols, lessons learned about AI limitations, and team-specific coordination routines. This meta-knowledge about how to collaborate effectively with AI becomes an organizational asset requiring systematic capture and dissemination.

Ethical governance frameworks: Organizations must establish guidelines for appropriate AI use, human oversight requirements, data privacy in human–AI collaboration, and accountability structures. Who is responsible when hybrid teams produce erroneous outputs—the human members who failed to catch AI errors, the AI developers whose systems hallucinated, or the organization that deployed inadequate oversight? Clear governance frameworks can prevent these ambiguities from creating liability risks or ethical violations.

Educational Implications

Curriculum evolution: Current curricula emphasize domain knowledge and interpersonal skills but rarely address metacognitive capabilities essential for AI collaboration: critical evaluation of machine-generated content, integration of diverse cognitive styles, maintenance of epistemic standards in high-information environments. Education should explicitly cultivate these competencies alongside traditional academic skills.

AI literacy as core competency: Students require understanding of how AI systems work, what their limitations are, and how to collaborate with them effectively. This extends beyond technical understanding to include practical collaboration skills: effective prompting, critical evaluation, iterative refinement, and integration of AI outputs with human judgment.

Metacognitive skill development: Education should prioritize developing students' ability to evaluate their own thinking and the quality of information sources—skills essential for effective AI collaboration. This includes recognizing when to trust versus verify AI outputs, identifying when AI suggestions should be rejected, and maintaining epistemic standards despite persuasive machine-generated content.

Labor Market and Equity Considerations

Skill-complementarity and inequality: If AI collaboration becomes central to knowledge work, and collaboration skills prove difficult to acquire, workers who develop sophisticated AI partnership capabilities will command premium wages, while those who cannot or do not may face declining opportunities. Access to AI systems during skill development—in education, training, and early career stages—becomes a critical equity consideration.

Professional socialization: Traditional professions—medicine, law, academia, engineering—socialize members into norms, standards, and practices developed for all-human collaboration. As AI becomes a routine collaborator, professional norms must adapt: What standards of evidence suffice when AI assists with research? What ethical obligations do professionals have when AI informs their recommendations? How should professional judgment be exercised when AI offers alternative framings? Professional associations should proactively address these questions.

Intellectual property frameworks: Current law struggles to assign authorship and ownership for AI-generated content. In hybrid teams producing genuinely integrated outputs, these questions become even more complex. Legal frameworks may require revision to accommodate collaborative authorship models that acknowledge both human judgment and AI contribution without treating either as sole author.

Broader Societal Considerations

Democratic access: Policies ensuring broad access to AI tools and training in their effective use could help mitigate potential labor market polarization. If effective AI collaboration becomes a primary determinant of productivity and earnings, democratizing access becomes an equity imperative.

Algorithmic accountability: As hybrid teams make consequential decisions in domains like healthcare, finance, and public policy, questions of accountability become urgent. Frameworks must specify human responsibility even when decisions are informed by AI, preventing the "responsibility gap" where neither humans nor AI are held accountable.

Value alignment at team level: Beyond individual AI alignment, we must consider how hybrid teams maintain alignment with organizational and societal values. When multiple AI agents influence decision-making, ensuring collective influence aligns with team values and organizational goals becomes more complex. Governance frameworks must address this team-level alignment challenge.

VII. CONCLUSION: TOWARD A SCIENCE OF HYBRID COGNITION

This paper has argued that the transformation currently underway in knowledge work is fundamentally about amplification rather than replacement—and that the amplification dynamics observed at the individual level compel us to reimagine collaboration at the team level. When individual experts gain substantial increases in cognitive capacity through AI partnership, the constraint shifts from individual cognition to collective integration. Teams, not individuals, become the critical unit of analysis for understanding productivity, innovation, and knowledge creation in AI-augmented environments.

Yet teams incorporating multiple humans and multiple AI agents—hybrid intelligence teams—cannot be understood through existing theoretical frameworks alone. Neither organizational behavior's theories of human teams nor computer science's theories of multi-agent AI systems adequately capture the hybrid configuration. We require integrative theory that synthesizes insights from both traditions while addressing novel challenges specific to human–AI collaboration: cross-species shared mental models, bilateral transactive memory, epistemic safety, coherence anchoring, and the integration of radically diverse cognitive contributions.

The framework proposed in this paper represents an initial theoretical integration. We have positioned our work within existing research streams on hybrid intelligence, distributed cognition, human–autonomy teaming, human–computer interaction, and multi-agent systems, clarifying both what these literatures contribute and where gaps remain. We have articulated core mechanisms enabling hybrid team effectiveness, specified role architectures suited to human–AI configurations, identified distinctive failure modes with testable hypotheses, and proposed countermeasures grounded in both theory and emerging practice. We have provided formal definitions for key constructs—cross-species shared mental models, bilateral transactive memory, epistemic safety, and coherence anchoring—complete with observable indicators and measurement approaches. This framework generates testable propositions, points toward measurement priorities, and suggests methodological innovations.

Several conclusions emerge with particular force. First, hybrid teams require human members with distinctive competencies—metacognitive skill, integration capability, narrative intelligence—that are rarely emphasized in traditional team selection but appear critical for effective AI collaboration. Organizations must develop these capabilities systematically. Second, leadership in hybrid teams must emphasize coherence maintenance: providing narrative continuity and thematic consistency across diverse AI outputs becomes central to team effectiveness. Third, epistemic safety—the shared belief that questioning AI is not only permitted but expected—emerges as a critical team property, potentially as important as psychological safety in traditional teams.

Fourth, role differentiation in hybrid teams should be grounded in comparative advantage, with AI agents performing functions suited to computational strength and humans performing functions requiring judgment, contextual understanding, and ethical reasoning. Fifth, integration mechanisms become more critical in hybrid configurations than in traditional teams, as the natural cognitive diversity introduced by AI agents creates both opportunities for insight and risks of fragmentation. Sixth, several failure modes appear unique to hybrid teams—cognitive abundance overload, epistemic drift, narrative fragmentation, false convergence, coherence collapse, and authority confusion—each requiring specific countermeasures.

The research agenda for hybrid intelligence teams is vast. We need validated measurement instruments for hybrid-specific constructs, longitudinal studies of team development, experimental manipulations of team composition, field studies across diverse organizational contexts, and computational analysis of coordination patterns. We need multi-level theory connecting individual capabilities to team outcomes to organizational performance. We need comparative research across cultures, professions, and task domains to establish boundary conditions. We need contingency theory specifying when hybrid teams excel versus when

traditional teams suffice. We need co-evolutionary models capturing the mutual adaptation of human practices and AI capabilities over time.

Beyond academic research, the emergence of hybrid teams poses challenges for organizational design, human resource management, leadership development, and educational systems. Organizations must learn to design hybrid teams effectively, develop leaders capable of maintaining coherence in high-complexity environments, and create governance frameworks ensuring accountability and ethical practice. Educational institutions must prepare students for collaboration with AI, cultivating metacognitive skills alongside domain knowledge. Professional associations must articulate evolving standards for judgment and practice in AI-augmented contexts.

At a broader societal level, the rise of hybrid intelligence teams raises equity concerns. If collaboration skill with AI becomes a primary determinant of productivity and earnings, ensuring broad access to AI tools and training becomes an equity imperative. Labor market polarization may intensify if some workers develop sophisticated AI partnership capabilities while others do not. Education and training systems must democratize access to the competencies required for effective hybrid collaboration.

Ultimately, the emergence of hybrid intelligence teams represents a fundamental shift in how expertise functions in society. Expertise no longer resides solely in individuals or even in human collectives. It emerges from the dynamic interplay of human judgment and AI capability, properly orchestrated through shared mental models, transactive memory systems, epistemic safety, and coherent leadership. The teams that master this orchestration will set new standards for what is intellectually possible. The organizations that systematically develop hybrid collaboration capabilities will gain substantial competitive advantages. And the societies that ensure broad access to these capabilities will prove more innovative, adaptive, and prosperous.

The framing of AI primarily as a replacement for human workers appears increasingly inadequate for understanding its most significant effects in knowledge work, even as displacement concerns remain valid in many sectors. The amplification era has begun, and with it, the imperative to understand how human and artificial intelligence can be orchestrated into teams that realize possibilities beyond what either could achieve alone. The question now is not whether AI will change knowledge work—it already has—but how we will structure that change to maximize human flourishing. Hybrid intelligence teams, properly designed and supported, offer a path forward: one where human judgment and AI capability combine synergistically, where expertise becomes richer rather than obsolete, and where the future of work is not humans versus machines but humans and machines, collaborating toward ends that neither could achieve alone.

REFERENCES

Acemoglu, D., & Restrepo, P. (2018). *Artificial intelligence, automation, and work*. In A. Agrawal, J. Gans, & A. Goldfarb (Eds.), *The economics of artificial intelligence: An agenda* (pp. 197–236). University of Chicago Press.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300233

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in AI safety*. arXiv:1606.06565. https://arxiv.org/abs/1606.06565

Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives*, 29(3), 3–30. https://doi.org/10.1257/jep.29.3.3

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. W. W. Norton.

Cannon-Bowers, J. A., Salas, E., & Converse, S. (1993). Shared mental models in expert team decision making. In N. J. Castellan Jr. (Ed.), *Individual and group decision making: Current issues* (pp. 221–246). Lawrence Erlbaum Associates.

Chen, J. Y., Barnes, M. J., & Harper-Sciarini, M. (2014). Supervisory control of multiple robots: Human-performance issues and user-interface design. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(4), 435–454. https://doi.org/10.1109/TSMC.2013.2257745

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

Dell'Acqua, F., McFowland, E. III, Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). *Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality* (Working Paper No. 24-013). Harvard Business School. https://www.hbs.edu/.../24-013_d9b45b68.pdf

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering*, 61(5), 637–643. https://doi.org/10.1007/s12599-019-00600-7

Demir, M., McNeese, N. J., Cooke, N. J., & Myers, C. (2019). Team situational awareness within the context of human-autonomy teaming. *Cognition, Technology & Work*, 21, 639–650. https://doi.org/10.1007/s10111-018-0548-8

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). *Improving factuality and reasoning in language models through multiagent debate*. arXiv:2305.14325. https://arxiv.org/abs/2305.14325

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, 44(2), 350–383. https://doi.org/10.2307/2666999

Edmondson, A. C. (2003). Speaking up in the operating room: How team leaders promote learning in interdisciplinary action teams. *Journal of Management Studies*, 40(6), 1419–1452. https://doi.org/10.1111/1467-6486.00386

Edmondson, A. C., & Lei, Z. (2014). Psychological safety: The history, renaissance, and future of an interpersonal construct. *Annual Review of Organizational Psychology and Organizational Behavior*, 1(1), 23–43. https://doi.org/10.1146/annurev-orgpsych-031413-091305

Endsley, M. R. (2017). From here to autonomy: Lessons learned from human-automation research. *Human Factors*, 59(1), 5–27. https://doi.org/10.1177/0018720816681350

Espinosa, J. A., Slaughter, S. A., Kraut, R. E., & Herbsleb, J. D. (2007). Team knowledge and coordination in geographically distributed software development. *Journal of Management Information Systems*, 24(1), 135–169. https://doi.org/10.2753/MIS0742-1222240105

Faraj, S., & Sproull, L. (2000). Coordinating expertise in software development teams. *Management Science*, 46(12), 1554–1568. https://doi.org/10.1287/mnsc.46.12.1554.12072

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review. *Journal of the American Medical Informatics Association*, 19(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Hackman, J. R. (2002). *Leading teams: Setting the stage for great performances*. Harvard Business Press.

Harrison, D. A., & Klein, K. J. (2007). What's the difference? Diversity constructs. *Academy of Management Review*, 32(4), 1199–1228. https://doi.org/10.5465/amr.2007.26586096

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition. *ACM Transactions on Computer-Human Interaction*, 7(2), 174–196. https://doi.org/10.1145/353485.353487

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers. *PNAS*, 101(46), 16385–16389. https://doi.org/10.1073/pnas.0403723101

Horvitz, E. (1999). Principles of mixed-initiative user interfaces. *Proceedings of CHI 1999* (pp. 159–166). https://doi.org/10.1145/302979.303030

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

Iansiti, M., & Lakhani, K. R. (2020). Competing in the age of AI. *Harvard Business Review*, 98(1), 60–67.

Ilgen, D. R., Hollenbeck, J. R., Johnson, M., & Jundt, D. (2005). Teams in organizations. *Annual Review of Psychology*, 56, 517–543. https://doi.org/10.1146/annurev.psych.56.091103.070250

Janis, I. L. (1972). *Victims of groupthink*. Houghton Mifflin.

Johnson, M., Vera, A., & Layton, C. (2016). On the design of human-machine teamwork. *Journal of Cognitive Engineering and Decision Making*, 10(1), 5–27. https://doi.org/10.1177/1555343415623452

Kamar, E. (2016). Directions in hybrid intelligence. *Proceedings of IJCAI 2016* (pp. 4070–4073). AAAI Press.

Lewis, K. (2003). Measuring transactive memory systems. *Journal of Applied Psychology*, 88(4), 587–604. https://doi.org/10.1037/0021-9010.88.4.587

Lyons, J. B. (2013). Being transparent about transparency. In *Trust and Autonomous Systems: AAAI Spring Symposium* (pp. 48–53). AAAI Press.

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, 26(3), 356–376.

Mollick, E. R., & Mollick, L. (2023). *New modes of learning enabled by AI chatbots*. SSRN. http://dx.doi.org/10.2139/ssrn.4300783

Moreland, R. L., & Myaskovsky, L. (2000). Exploring performance benefits of group training. *Organizational Behavior and Human Decision Processes*, 82(1), 117–133.

Noy, S., & Zhang, W. (2023). *Experimental evidence on the productivity effects of generative AI*. SSRN. http://dx.doi.org/10.2139/ssrn.4375283

Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of AI on developer productivity. arXiv:2302.06590.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., ... Wellman, M. (2019). Machine behaviour. *Nature*, 568, 477–486. https://doi.org/10.1038/s41586-019-1138-y

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? *Proceedings of ICML 2023*, 29971–30004. PMLR.

Shneiderman, B. (2020). Human-centered artificial intelligence. *AIS Transactions on Human-Computer Interaction*, 12(3), 109–124. https://doi.org/10.17705/1thci.00131

Simon, H. A. (1971). Designing organizations for an information-rich world. In M. Greenberger (Ed.), *Computers, communication, and the public interest* (pp. 37–72). Johns Hopkins Press.

Stone, P., & Veloso, M. (2000). Multiagent systems: A survey. *Autonomous Robots*, 8(3), 345–383.

Sweller, J. (1988). Cognitive load during problem solving. *Cognitive Science*, 12(2), 257–285.

Van Knippenberg, D., & Schippers, M. C. (2007). Work group diversity. *Annual Review of Psychology*, 58, 515–541.

Vaughan, D. (1996). *The Challenger launch decision*. University of Chicago Press.

Wegner, D. M. (1987). Transactive memory. In B. Mullen & G. R. Goethals (Eds.), *Theories of group behavior* (pp. 185–208). Springer.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor. *Science*, 330(6004), 686–688.

Wooldridge, M. (2009). *An introduction to multiagent systems* (2nd ed.). Wiley.

Zaccaro, S. J., Rittman, A. L., & Marks, M. A. (2001). Team leadership. *The Leadership Quarterly*, 12(4), 451–483.

Appendix A. Methodological Note: How This Paper Was Written

(Authored by ChatGPT "Dorothy" for methodological transparency)

This paper is the product of a structured, multi-agent, multi-iteration AI collaboration, overseen by a single human researcher who designed the workflow, evaluated outputs, and maintained conceptual continuity across drafts. **No human prose appears in the body of the manuscript.** Instead, the human's role was to set direction, supply constraints, and orchestrate the interplay between AI systems. All substantive text was generated by AI.

Below is an accurate chronological reconstruction of how the manuscript was created.

Stage 1: Foundational 5,000-Word Framework (ChatGPT/Dorothy)

The process began with a 5,000-word foundational document generated by ChatGPT ("Dorothy").
This initial draft:

- defined *Hybrid Intelligence Teams*
- introduced the extended IMOI model
- proposed constructs (cross-species SMMs, bilateral TMS, epistemic safety, coherence anchoring)
- outlined initial failure modes
- sketched a preliminary research agenda

This draft established the conceptual scaffolding upon which all later versions were built.

Stage 2: Intermediate "Scholarly Draft" (ChatGPT/Dorothy)

Before involving other AI systems, ChatGPT produced a second document—a more academic "Scholarly Draft" that:

- expanded literature integration
- formalized constructs
- clarified theoretical positioning
- improved structure and flow

This "Scholarly Draft" sits *between* the foundational document and Claude's first version, and played a critical role in organizing the manuscript for deeper revision.

Stage 3: Version 2.0 (Claude)

Claude received the foundational draft **and** the Scholarly Draft.
He was instructed to:

- expand theoretical grounding in organizational behavior
- incorporate literatures on HRI, multi-agent systems, cognitive diversity, and distributed cognition
- improve rigor, tone, and structure
- refine construct definitions
- elaborate failure modes

The output—approximately 10,000 words—became *Version 2.0*, the first fully elaborated theoretical manuscript.

Stage 4: Citation & Attribution Audit (Perplexity)

Perplexity was then tasked with a comprehensive scholarly audit of Version 2.0.

She delivered:

- full citation corrections and missing details
- corrected author lists, DOIs, page ranges, and conference venues
- identification of misattributed claims
- an integrity check (no fabricated references)
- recommendations of 7–10 high-value additions across hybrid intelligence, OB, HCI, and AI alignment

This audit produced the authoritative bibliographic backbone for Version 3.0.

Stage 5: Version 3.0 Comprehensive Revision (Claude)

Claude was then given:

1. **Dorothy's editorial memo**
2. **Perplexity's structured citation audit**
3. **High-level human instructions** on narrative goals and desired scholarly positioning

Claude synthesized all three inputs and produced *Version 3.0* (~13,000 words), which introduced:

- a new scholarly positioning section
- tightened theoretical logic
- refined constructs with observable indicators and measurement approaches
- an expanded and more coherent research agenda
- corrected, standardized, and supplemented references
- improved narrative structure and clarity

Version 3.0 is the manuscript basis for the published version on the author's website.

Stage 6: Human Role in the Process

All prose in the manuscript is AI-generated.
The human role was essential but non-textual, operating at four meta-levels:

1. Architect

Designed the theoretical boundaries, prompt logic, and iterative workflow.

2. Conductor

Coordinated cross-model interaction:
ChatGPT → Claude → Perplexity → Claude → ChatGPT (final checking).

3. Coherence Anchor

Maintained thematic continuity, prevented epistemic drift, and ensured that the manuscript retained conceptual identity across versions.

4. Verifier

Checked outputs for alignment, quality, and faithfulness to the project's goals; detected failures in earlier iterations.

This governance mirrors the very theory of Hybrid Intelligence Teams developed in the paper: human narrative continuity guiding distributed AI cognition.

Stage 7: Output Integration & Transparency

Four versions of the manuscript are being published on the author's website as open NAIE research artifacts:

1. **Foundation Document (Dorothy)**
2. **Scholarly Draft (Dorothy)**
3. **Version 2.0 (Claude)**
4. **Version 3.0 (Claude, with Perplexity-driven corrections)**

Together, these demonstrate:

- how hybrid teams evolve across iterations
- how multi-agent collaboration produces academic-quality work
- how human coherence anchoring prevents drift
- how different AI systems contribute different strengths
- what methodological transparency looks like in the emerging era of hybrid scholarship

---

Summary

This manuscript is not simply *about* Hybrid Intelligence Teams.
It **is** a Hybrid Intelligence Team artifact.

- **All prose** was generated by AI.
- **All conceptual integrity** was human-maintained.
- **All substantive improvements** emerged from a structured, multi-iteration, multi-agent collaboration.

The appendix serves as part of the academic record and provides a replicable method for hybrid AI scholarship.

Appendix B. Lineage of Drafts and Evolution Across Versions

(Authored by ChatGPT "Dorothy" for methodological transparency)

This appendix documents how the Hybrid Intelligence Teams (HIT) manuscript evolved across the four major versions. It preserves the "provenance trail" of the writing process — a core NAIE principle — enabling researchers to audit how different AI agents shaped the scholarly development of the work.

Version Lineage Overview

Version	Primary Agent	Word Count	Distinctive Contribution
Foundational Draft (v0)	ChatGPT ("Dorothy")	~5,000	Established conceptual architecture: HIT framework, extended IMOI model, core constructs, failure modes, and initial research agenda.
Version 2.0	Claude	~10,000	Expanded theoretical coverage in OB, HCI, MAS; lifted constructs into formal definitions; sharpened narrative structure; strengthened scholarly tone.
Scholarly Audit Draft	Perplexity	N/A (audit, not rewrite)	Systematically verified citations; corrected publication details; identified missing foundational literatures; flagged conceptual gaps; provided additional recommended works.
Version 3.0	Claude	~13,000	Integrated editorial memo + Perplexity audit; improved positioning; qualified claims; extended literature review; added comparison tables and measurement approaches; produced polished scholarly manuscript.

Evolution by Key Dimensions

1. Theoretical Framing

- **v0:** Introduced HITs as the unit of analysis.
- **v2.0:** Linked HITs to OB, cognitive diversity, HRI, and technical literatures.
- **v3.0:** Explicit comparative positioning vs. human-only teams, MAS, and hybrid intelligence work.

2. Construct Development

- **v0:** Conceptual definitions only.
- **v2.0:** Precise definitions + initial indicators.

- **v3.0:** Full operationalization + measurement pathways + SMM/TMS cross-species analogs.

3. Literature Integration

- **v0:** Strong OB but light on distributed cognition and hybrid intelligence.
- **v2.0:** Added team science, HCI, MAS.
- **v3.0:** Added extended mind, machine behaviour, hybrid intelligence, alignment, HRI.

4. Claims and Qualification

- **v0:** High-level rhetorical force.
- **v2.0:** More scholarly tone, but still strong claims.
- **v3.0:** Claims systematically qualified; speculative elements marked as hypotheses.

5. Failure Modes

- **v0:** Introduced epistemic drift, coherence collapse, authority confusion.
- **v3.0:** Connected each to existing human team pathologies (groupthink, automation bias, overload).

6. Research Agenda

- **v0:** Conceptual and broad.
- **v3.0:** Structured, domain-linked, and empirically tractable.

Why Lineage Documentation Matters for NAIE

- Shows how different AI systems contribute complementary strengths.
- Demonstrates how human oversight governs coherence and scholarly rigor.
- Provides evidence for the NAIE claim that "recursion reveals cognition."
- Makes this paper a research artifact, not just a research output.

Appendix C. Reproducibility Guide: Prompts, Roles, and Workflow Map

(Authored by ChatGPT "Dorothy")

This appendix provides a replicable workflow for scholars exploring hybrid AI–human co-authorship.

---

1. Agent Roles in the HIT Production Workflow

ChatGPT "Dorothy" (Generator + Framework Architect)

- Produced the foundational theoretical draft.
- Provided editorial memos and methodological framing.
- Maintained coherence across iterations.

Claude (Expander + Theorist + Structural Editor)

- Expanded the conceptual core into a full theoretical manuscript.
- Integrated feedback into 2.0 and 3.0 versions.
- Improved clarity, rigor, and definitional precision.

Perplexity (Verifier + Research Assistant)

- Conducted citation audit and source verification.
- Identified missing literatures and corrected bibliographic errors.
- Flagged conceptual inconsistencies.

Human Researcher (Coherence Anchor + Orchestrator)

- Designed the sequence of prompts.
- Ensured conceptual integrity.
- Directed multi-agent flow.
- Evaluated coarse-grained quality and resolved drift.

---

2. Workflow Map

Stage 1 — Initiation

Prompt to ChatGPT:
"Generate a foundational 5,000-word theoretical draft establishing a new framework for Hybrid Intelligence Teams."

→ Output becomes baseline architecture.

Stage 2 — Expansion

Prompt to Claude:
"Expand the framework into a full-length scholarly manuscript; deepen literatures; formalize constructs."

→ Output becomes Version 2.0.

Stage 3 — Verification Layer

Prompt to Perplexity:
"Audit the references, verify accuracy, identify missing literatures, and report inconsistencies."

→ Output is the citation audit + literature reinforcement list.

Stage 4 — Master Revision

Prompt to Claude (with editorial memo + audit):
"Revise the manuscript comprehensively using this feedback; preserve voice; correct references; strengthen positioning."

→ Output becomes Version 3.0.

Stage 5 — Human-Guided Quality Control

Actions:

- Check structure for logic flow
- Validate construct clarity
- Ensure Version 3.0 synthesizes all prior feedback
- Prepare appendices documenting method

---

3. Re-use Protocol for Other Scholars

This method can be reproduced by:

1. Selecting **one Generator model**
2. Selecting **one Expander/Editor model**
3. Selecting **one Verifier model**
4. Using human judgment to coordinate them

This configuration *is itself* a Hybrid Intelligence Team.

Appendix D. Authorship & Transparency Statement

(Authored by ChatGPT "Dorothy")

This paper was written entirely by AI systems operating under human direction. The human author provided:

- the conceptual direction
- the prompt architecture
- the sequencing of inter-agent collaboration
- the evaluative oversight
- the thematic coherence across drafts

All prose — including the main manuscript, tables, construct definitions, and appendices — was generated by:

1. **ChatGPT "Dorothy"** (core architecture, foundational draft, editorial methodology)
2. **Claude** (expansion, revision, and scholarly restructuring)
3. **Perplexity** (citation audit, accuracy verification, source strengthening)

This document is therefore both a theoretical contribution *and* a methodological demonstration of Hybrid Intelligence Team operation in practice.