FROM LOOP TO PARTNGRSHIP: A FRAMGWORK FIOR UNDGRSTANDING THG EVOLVING PARADIGMS OFI HUMAN- AI COLLABORATION

Sagar Patil
DBA18, SP Jain School of Global Management
sagar.ds22dba016@spjain.org
Suneel Sharma
Principal Supervisor
sunil2500@gmail.com
Mehregan Mahdav
Co-Supervisori
mehregan_m@hotmail.com
DOI: https://doi.org/10.64558/bllp/930827krohkn

ABSTRACT

The term 'Human-in-the-Loop' (HIL), once a clarifying concept for AI systems reliant on human input, now obscures more than it reveals. Its monolithic application to a vast and expanding range of human-AI interactions has led to a conceptual fragmentation that impedes scientific progress and responsible systems' development. This research confronts this ambiguity by proposing a new, tripartite framework that offers a more granular and robust typology for understanding human-AI collaboration. This framework is the result of a systematic literature review designed to synthesize a fragmented, interdisciplinary body of research. We first establish a foundational dichotomy based on the locus of control and introduce a tripartite framework consisting of: HIL (AI-led), where AI is autonomous, and humans provide input for improvement (e.g., data labeling for self-driving cars);

AI2L (Human-led), where humans are in control, and AI augments their capabilities (e.g., AI- powered diagnostic tools for doctors); And Hybrid Intelligence (HI), a more advanced paradigm focused on symbiotic, co-creative partnerships between humans and AI, often enabled by generative AI (e.g., generative design, collaborative software engineering).

It demonstrates that the choice of the human-AI collaboration paradigm is strongly influenced by the domain's primary objectives: efficiency for HIL, accountability for AI2L, and creativity/innovation for HI. It also establishes that ethical considerations must move beyond simple human oversight ("human in the loop") to a more robust and systemic approach called "participatory governance," which involves diverse stakeholders throughout the AI's lifecycle to ensure fairness and accountability. This resolves a central paradox in human-AI studies,

re-conceptualizes the landscape of collaboration, and establishes a new foundation for the design and governance of socio-technical systems.

INTRODUCTION:

BEYOND A MONOLITHIC "HUMAN-IN-THE-LOOP" THE PERFORMANCE PARADOX: A FIELD IN CRISIS

While the promise of human-AI collaboration is immense, recent large-scale reviews reveal a startling paradox: on average, these hybrid teams fail to outperform AI acting alone (Malone et al., 2025). A comprehensive review of over 100 experiments found that, on average, human-AI combinations do not outperform the best-performing individual system, whether that is the AI alone or the human alone (Malone et al., 2025). For instance, in tasks like detecting fake hotel reviews, AI alone was more accurate than the human-AI combination (Malone et al., 2025). Our assumptions about synergy may be incorrect, indicating a need for better conceptual frameworks to understand the dynamics of effective collaboration.

The varied conceptualizations of human-AI interaction extend beyond mere semantics, indicating a pivotal paradigm shift within computing. Initially, AI emphasized classification and prediction, relegating human involvement to data labeling and verification. This aligns with the foundational Human-in-the-Loop (HIL) paradigm, wherein humans oversee or offer authoritative input to the machine (Arambepola & Munasinghe, 2021; Bastani et al., 2017; Natarajan et al., 2024). Within this framework, the AI produces a distinct, verifiable output (e.g., a "cat" label or a score), and the human's function is to furnish or validate it.

Consequently, the "loop" delineates a direct supervisory relationship.

The emergence of generative artificial intelligence (AI), capable of producing novel and intricate outputs such as text, imagery, and programming code, has revealed new opportunities for collaborative creation and partnership (Rafner et al., 2024). Unlike conventional AI, which primarily yields definitively "correct" or "incorrect" outcomes, generative AI provides initial frameworks suitable for interpretation, refinement, and iterative development. This paradigm shift from classification to generation fundamentally alters the essence of AI's output and, consequently, the human role. Individuals transcend a merely supervisory capacity to become integral collaborators and partners within a constructive process. The continued application of the antiquated term "HIL" (Human-in-the-Loop) to characterize this novel and distinct relationship underlies a conceptual dilemma. The discourse is progressing from an emphasis on feedback loop mechanisms toward a more sophisticated examination of agency, cooperation, and governance (Arambepola & Munasinghe, 2021).

Research Questions

Primary Research Question (PRQ)

Proposed Question:

How can the landscape of human-AI collaboration be re-conceptualized beyond the monolithic and ambiguous 'Human-in-the-Loop' (HIL) paradigm to resolve the field's "performance

paradox" and provide a more granular framework for analyzing, designing, and governing socio- technical systems?

Justification: This central inquiry animates the entire research by addressing a foundational, dual crisis currently arresting the field of human-AI studies. The first is an empirical **"performance paradox"**: the persistent and counterintuitive failure of human-AI teams, on average, to outperform AI acting alone. It argues that this empirical anomaly is a direct symptom of a more profound, more corrosive **conceptual crisis**: the monolithic and ambiguous application of the term 'Human-in-the-Loop' (HIL), which now "obscures more than it reveals". This terminological confusion has precipitated a "conceptual fragmentation that impedes scientific progress and responsible systems development".

Therefore, this research question functions as a strategic intervention. It frames the entire project as a necessary response to this field-defining impasse, mandating the development of a new theoretical apparatus capable of resolving the paradox and providing the analytical precision required to rigorously design, evaluate, and govern the next generation of socio-technical systems.

Secondary Research Questions (SRQs)

The primary question is addressed through a series of more specific, answerable secondary questions that collectively structure the argument.

SRQ1: The Paradigms of Collaboration

Proposed Question: What are the defining characteristics, core principles, and critical distinctions of the primary paradigms of human-AI collaboration—specifically, AI-led supervision (HIL), human-led augmentation (AI-in-the-Loop, or AI2L), and symbiotic co- creation (Hybrid Intelligence)?

Justification: This question erects the theoretical architecture of the entire research. It is the primary mechanism for resolving the field's "conceptual crisis." It requires the systematic deconstruction of the ambiguous 'Human-in-the-Loop' monolith and the subsequent construction of a more precise and powerful typology. The central contribution is this new framework, which is built upon two critical pillars.

First, it establishes the foundational dichotomy between HIL (AI-led, automation-focused) and AI2L (human-led, augmentation-focused)—a distinction mandated by a growing body of scholarship that critiques the misapplication of the HIL term and demands analytical clarity based on the locus of control. Second, it introduces Hybrid Intelligence (HI) as a distinct, synthetic paradigm essential for conceptualizing the symbiotic, co-creative partnerships catalyzed by the advent of generative AI.

The answer to this question is ultimately codified in the comparative framework (Table 1),

which functions as the core analytical engine that the subsequent sections validate against real-world domains and extend into the ethical frontier.

SRQ2: The Paradigm-Domain Fit

Proposed Question: How does the choice of a human-AI collaboration paradigm correlate with the primary value drivers of different application

domains, such as efficiency, accountability, and creativity?

Justification: This question drives the empirical grounding and validation of the entire theoretical apparatus. Having established the distinct collaboration paradigms, testing their utility and explanatory power against real-world phenomena is essential. This question posits and seeks to substantiate the **"paradigm-domain fit"** hypothesis—a core proposition asserting that the selection of a collaborative model is not arbitrary but is instead strongly predicted by a domain's fundamental value drivers.

To answer this, the research marshals' evidence from a cross-domain analysis, hypothesizing that high-stakes domains like healthcare and finance, which are governed by the imperative of **accountability**, will gravitate towards the human-led **AI2L** paradigm. Conversely, domains driven by the relentless pursuit of scale and **efficiency**, such as social media content moderation and autonomous systems development, will necessarily rely on the AI-led **HIL** model. Finally, it predicts that fields where the primary currency is **creativity** and novel ideation, like generative design and software engineering, are the natural incubators for the symbiotic **HI** paradigm. By answering this question, the research moves beyond theoretical classification to demonstrate the framework's capacity to explain and predict observable patterns of AI adoption across the global economy.

SRQ3: The Evolution of Ethical Oversight

Proposed Question: As human-AI collaboration matures from simple supervision to complex partnership, how must the corresponding ethical framework evolve beyond the simplistic notion of "human in the loop" to ensure robust fairness, accountability, and alignment with societal values?

Justification: This question propels the research into its crucial normative dimension, extending the analysis from the descriptive and explanatory to the prescriptive. It directly confronts the ethical insufficiency of the simplistic "human in the loop" safeguard, which the research argues has devolved into a fragile and often misleading concept. This simplistic backstop can create "collaboration theatre" or "participation washing," providing a veneer of legitimacy without genuine oversight and often designating a "scapegoat-in-the-loop" to bear responsibility for systemic failures.

Answering this question is essential, as it posits that the ethical framework must co-evolve

with the collaborative paradigms. The nature of risk shifts—from biased human input in HIL systems, to automation bias in AI2L, and to diffuse accountability in HI partnerships demands a more sophisticated ethical response than a single, static solution. Ultimately, this inquiry sets the stage for a most forward-looking contribution: the proposal of "Participatory Governance" as a more robust, systemic model that expands the "loop" to include a diverse range of stakeholders in a structured, ongoing process, ensuring that complex AI systems are not only practical but also just and accountable.

Research Objectives

Primary Research Objgctivg

To resolve the conceptual ambiguity surrounding 'Human-in-the-Loop' by developing and substantiating a new tripartite framework for human-AI collaboration (HIL, AI2L, and HI) grounded in a systematic synthesis of interdisciplinary literature.

Justification: This objective is the central pillar of the entire research endeavor, corresponding directly to the Primary Research Question. It encapsulates the principal scholarly contribution: the development and substantiation of a novel analytical framework. This is not merely a classification exercise; it is a direct intervention designed to resolve the "conceptual crisis" and terminological ambiguity that currently plagues the field of human-AI studies. Crucially, the objective specifies the rigorous methodology employed—a systematic synthesis of a fragmented, interdisciplinary body of literature, grounding the new framework in a comprehensive evidence base. By framing the goal in this manner, it positions itself not as a passive review of existing work but as an act of theory-building that furnishes the field with a new, more powerful analytical engine for understanding and shaping human-AI collaboration.

Secondary Research Objectives

The primary objective is achieved by completing several secondary objectives, each aligned with a secondary research question.

Objective 1: To systematically define and differentiate the HIL, AI2L, and HI paradigms based on the locus of control, primary goals, human/AI roles, core challenges, and evaluation metrics, codifying these distinctions in a comparative framework (Table 1).

Justification: This objective operationalizes the first secondary research question and represents the core constructive act. It moves beyond high-level critique to the granular, definitional work required to build the new typology. Systematically dissecting the paradigms across multiple dimensions—from the locus of control to evaluation metrics—provides the intellectual rigor necessary for the framework to be a credible analytical tool. Crucially, this objective culminates in creating Table 1, a key scholarly artifact that codifies the research's central theoretical contribution into a concise, accessible, and citable format for researchers and practitioners alike.

Objective 2: To analyze and provide evidence for the "paradigm-domain fit" hypothesis by examining case studies of human-AI collaboration in domains driven by accountability (e.g., healthcare, finance), efficiency (e.g., content moderation, autonomous systems), and creativity (e.g., generative design, software engineering*)*.

Justification: This objective operationalizes the second secondary research question and empirically grounds the theoretical framework. It moves beyond abstract classification to demonstrate its real-world explanatory and predictive power. This is achieved by testing the central **"paradigm-domain fit"** hypothesis, which posits that

the choice of a collaborative model is not arbitrary but is instead strongly correlated with a domain's primary value drivers. By marshaling evidence from case studies across distinct sectors—linking the imperative of **accountability** to the **AI2L** model in healthcare, the drive for **efficiency** to the **HIL** model in content moderation, and the pursuit of **creativity** to the **HI** model in generative design—this objective validates the framework's utility as a powerful analytical tool for making sense of observable patterns of AI adoption in practice.

Objective 3: To critique the ethical insufficiency of the "human in the loop" safeguard and propose "Participatory Governance" as a more robust, systemic model for ensuring fairness and accountability in complex AI systems.

Justification: This objective operationalizes the third secondary research question and constitutes the normative and ethical centerpiece of the research. It moves the argument from the explanatory to the prescriptive by first launching a necessary critique of the "human in the loop" safeguard, which has proven to be a fragile and often misleading ethical backstop. The research argues this concept frequently devolves into "collaboration theatre" or "participation washing," where the performance of oversight provides a veneer of legitimacy without genuine power redistribution, or worse, creates a "scapegoat-in-the-loop" to absorb liability for systemic failures.

Thesis Statement and Proposed Framework

This research introduces a novel tripartite framework, moving beyond the simplistic 'loop' model to offer a nuanced spectrum of human-AI collaboration, differentiated by agency and goals. We demonstrate that this framework offers the necessary granularity to elucidate the conditions under which human-AI synergy is—and is not—achieved, providing a clear path for future research and system design. Our research proposes a new tripartite framework for analyzing and designing human-AI systems to clarify this evolving landscape. This framework categorizes the spectrum of interaction into three distinct and progressively sophisticated paradigms:

- **The HIL/AI2L Dichotomy:** This fundamental distinction classifies systems by their primary locus of control. **Human-in-the-Loop (HIL)** systems are AI-led and human-supervised, whereas **AI-in-the-Loop (AI2L)** systems are human-led and AI-augmented. Researchers like Natarajan et al. (2024) advocate for this clear analytical lens, moving beyond ambiguity by focusing on decision-making authority.

- **Hybrid Intelligence (HI):** This advanced collaborative paradigm moves beyond a simple control dynamic to emphasize human-AI symbiosis. It centers on mutual learning and the co-creation of solutions by human and AI agents, often facilitated by modern generative AI capabilities (Mayer et al., 2024; Rafner et al., 2024). This signifies a shift towards a more holistic, human-centric view of technology and work

(Mayer et al., 2024).

- **Participatory Governance:** This represents the ethical and practical culmination of the field's maturation. For complex, high-stakes systems, the "loop" must extend beyond the immediate user or developer. It requires including a broad range of stakeholders in a structured, ongoing governance process. Griffen and Owens (2024) argue that this ensures systems are effective but also fair, accountable, and aligned with societal values.

The Foundational Dichotomy: Human-in-the-Loop (HIL) versus AI-in-the-Loop (AI2L)

We propose a fundamental distinction rooted in the locus of control to address the current terminological confusion in the field. We identify two primary paradigms: the AI-led, automation-centric **Human-in-the-Loop (HIL)** and the human-led, augmentation-focused **AI- in-the-Loop (AI2L)** (Natarajan et al., 2024). A thorough review of existing literature indicates that the broad term "Human-in-the-Loop" often merges these two distinct modes of interaction (Natarajan et al., 2024). As Natarajan et al. (2024) contend, numerous systems currently labeled as HIL are, in reality, misclassified and more precisely fit the description of AI2L. Recognizing this distinction carries significant ramifications for system design, evaluation, and ethical considerations.

The Human-in-the-Loop (HIL) Paradigm: AI-Led Automation

In the Human-in-the-Loop (HIL) paradigm, the AI system maintains control over the primary process and retains decision-making authority (Natarajan et al., 2024). While fundamentally autonomous, the system strategically incorporates human input for specific tasks. These tasks include error correction, performance supervision, and addressing edge cases outside the AI's operational confidence (Bastani et al., 2017; Natarajan et al., 2024). The primary goal of HIL is to improve the AI model itself—to enhance its accuracy, robustness, and scope of automation by leveraging human intelligence as a refinement tool (Singhal et al., 2023).

In this model, the human plays a specific, often circumscribed, role as an "oracle," a "supervisor," or a "teacher" for the AI (Bastani et al., 2017; Natarajan et al., 2024). This includes the canonical task of data annotation, where humans label vast datasets to train supervised learning models, a process that is often the most significant bottleneck in deploying machine learning systems (Arambepola & Munasinghe, 2021; Savage, 2023). It also encompasses providing evaluative feedback on model outputs and directly correcting model errors to fine-tune performance (Natarajan et al., 2024; Singhal et al., 2023). The focus is squarely on making AI-led automation more effective, with the human inputs used to "guide" the model to a better optimum (Natarajan et al., 2024).

The Human-in-the-Loop (HIL) paradigm is commonly applied in areas demanding large-scale operations and high efficiency. For instance, in social media content moderation, AI systems independently identify potentially harmful content.

Human moderators then assess flagged items, rendering final decisions on complex or nuanced cases that necessitate profound contextual comprehension (Singhal et al., 2023). Likewise, the creation of autonomous vehicles employs HIL processes, wherein humans annotate vast quantities of sensor data.

This teaches the vehicle's perception systems to identify objects and navigate their environment (Jayapradha et al., 2024). These interactions often occur via crowdsourcing platforms, which transform human contributions into a series of distinct micro-tasks, prioritizing cost-effectiveness and productivity (Amershi et al., 2014; Savage, 2023).

The AI-in-the-Loop (AI2L) Paradigm: Human-Led Augmentation

The AI2L paradigm fundamentally differs, positioning the human expert at the core with complete control and final decision-making power (Natarajan et al., 2024). This system's purpose is not to substitute human abilities but to enhance them, thereby amplifying rather than displacing human capabilities (Natarajan et al., 2024). AI functions as an intelligent assistant, a perception aid, or a sophisticated decision-support tool, providing analyses, summaries, and suggestions to inform human judgment (Natarajan et al., 2024). The primary goal of AI2L is to enhance humans' performance, efficiency, and the quality of their decisions. The focus is on empowering the human user, with the overall system supporting a human-led workflow (Natarajan et al., 2024).

AI2L systems are prevalent in high-stakes professional domains where accountability and expert judgment are paramount (Natarajan et al., 2024). In healthcare, AI-powered clinical decision support systems (CDSS) analyze medical images or patient data to highlight potential areas of concern, but the human physician makes the final diagnosis and determines the course of treatment (Arambepola & Munasinghe, 2021; Kumar et al., 2024). In finance, an AI model might recommend portfolio adjustments based on market analysis, but the human trader provides feedback, incorporates their strategic knowledge, and executes the final trades (Arambepola & Munasinghe, 2021). Other examples include interactive tools that use AI to surface distant but relevant scientific papers to augment the creativity of a researcher, or systems that help software developers by suggesting code completions or identifying potential bugs, while the developer retains control over the final codebase (Kang et al., 2022; Medepalli, 2025).

A Comparative Framework of HIL and AI2L Paradigms

Table 1 provides a comparative summary of the HIL and AI2L paradigms across several key dimensions to codify this foundational distinction. This framework distills the complex and emerging discourse in the literature into a clear, concise, and citable format, making the research's central theoretical contribution immediately apparent and providing an accessible analytical tool for researchers and practitioners.

Table 1: A Comparative Framework of HIL and AI2L Paradigms, synthesized from Arambepola & Munasinghe (2021), Bastani et al. (2017), Kang et al. (2022), Natarajan et al. (2024), and Singhal et al. (2023).

Dimension	Human-in-the-Loop (HIL) Paradigm	AI-in-the-Loop (AI2L) Paradigm
Locus of Control	AI-Led: The AI system is autonomous and controls the primary process.	Human-Led: The human expert is at the center and retains final decision-making authority.
Primary Goal	Automation: To improve the AI model's performance, accuracy, and robustness.	Augmentation: To enhance the human expert's performance, efficiency, and decision quality.
Human Role	Supervisor/Oracle/Teacher: Provides labels, feedback, and corrections to the AI.	Decision-Maker/User: Uses AI-generated insights as a tool to inform their judgment.
AI Role	Autonomous Agent: Performs the task and requests human input for complex cases.	Intelligent Assistant/Tool: Provides analysis, suggestions, and decision support to humans.
Core Challenge	Human Input Quality: Managing human bias, inconsistency, and the cost/scalability of annotation.	Human-AI Interaction: Designing effective interfaces, ensuring explainability, and mitigating automation bias.

Primary Evaluation Metrics	AI-Centric: Accuracy, precision, recall, and F1-score of the model.	Human-Centric: Quality of final human decision, user trust, cognitive load, task completion time, and overall system outcome.
Example Application	Content moderation on social media; large-scale data annotation for autonomous driving.	Clinical decision support in medicine, creative assistance for scientific research, and portfolio management in finance.

Hybrid Intelligence: A Synthesis for Symbiotic Collaboration

The evolution of collaborative systems does not culminate in the HIL/AI2L distinction. Propelled by the advent of generative AI, a third paradigm is emerging: **Hybrid Intelligence (HI)**. This is not a mere midpoint but a synthetic paradigm that reconceptualizes the human- AI relationship as a symbiotic, co-creative partnership (Mayer et al., 2024; Rafner et al., 2024). In HI systems, human and AI agents are conceptualized as collaborators working on joint tasks, where the goal is to leverage the complementary strengths of each to achieve a level of performance that surpasses what either could accomplish alone (Mayer et al., 2024).

Core Principles of Hybrid Intelligence

Hybrid Intelligence is characterized by a set of core principles that distinguish it from simpler loop-based models:

- **Co-Construction: A Collaborative Genesis of Solutions**

Central to HI collaboration is the process of **co-construction**, wherein both human and AI partners actively and iteratively engage in developing solutions and articulating problem objectives. This is not a unidirectional process of command and execution; rather, it is a dynamic interplay. Humans typically initiate the process by conveying high- level goals, preferences, constraints, and contextual intricacies that an AI might find challenging to infer independently. The AI then employs its computational power, data processing capabilities, and learned models to generate potential solutions, initial drafts, or conceptual pathways. These AI-generated outputs serve as tangible points for discussion and refinement by the human partner. Interaction frequently occurs through natural language communication, facilitating a more intuitive and flexible exchange of ideas. As noted by Dutta et al. (2024), this continuous cycle of human input, AI generation, and subsequent human critique and refinement ensures that the final solution is technically sound and consonant with human values, intentions, and intricate situational understanding. This iterative back-and-forth ensures the collaborative evolution of the solution, integrating the strengths of both intelligences.

- **Mutual Learning and Adaptation: A Bidirectional Growth Trajectory**

Beyond the simple exchange of information, HI collaboration cultivates a profound environment of **mutual learning and adaptation**. This constitutes a bidirectional learning process where both entities evolve their understanding and capabilities. The AI learns from human feedback, which is a critical data source for enhancing its models, comprehending human preferences, identifying biases, and refining its problem-solving strategies. This learning can be explicit, through direct corrections or ratings, or implicit, by observing human modifications to its outputs. Concurrently, humans learn from the AI's insights, including identifying patterns, exploring novel possibilities, or processing vast quantities of information beyond human cognitive capacity. AI can present alternative perspectives or efficient solutions to humans that might not have occurred,

thereby expanding their mental models and refining their workflows. This "hybrid collaborative learning cycle," as highlighted by Wiethof & Bittner (2022) and Rafner et al. (2024), leads to continuous improvement in both individual and collective performance. It fosters a shared understanding of the task, the solution space, and each other's capabilities, particularly in systems meticulously engineered to augment rather than supplant human-AI interaction. This continuous feedback loop propels an evolving partnership wherein each learns to leverage the other's strengths more effectively.

- **Shared and Dynamic Agency: Fluid Control for Adaptive Performance**

A distinguishing characteristic of effective HI collaboration is **shared and dynamic agency**. In contrast to traditional human-computer systems, where control is often fixed (e.g., the human user explicitly commands the software), HI systems demonstrate a fluid and dynamically allocated control distribution. This implies that leadership over specific sub-tasks can transition between the human and the AI based on the nature of the task, the requisite expertise, and the evolving demands of the situation. For instance, an AI might assume leadership in computationally intensive sub-tasks such as generating extensive code, executing complex simulations, analyzing extensive datasets, or predicting outcomes based on patterns. Conversely, humans consistently retain primary agency and control over strategic direction, the definition of creative objectives, the exercise of ethical judgment, the application of nuanced contextual understanding, and the final evaluation and decision-making. As Mayer et al. (2024) articulated, this flexible sharing of control enables the human-AI team to be highly adaptive, responding efficiently to evolving task demands and unforeseen challenges. This dynamic agency ensures that the strengths of human intuition, creativity, and ethical reasoning, as well as AI's speed, precision, and data processing power, are optimally leveraged throughout the collaborative process, leading to more robust and comprehensive outcomes.

Framgworks and Modgls fior Opgrationalizing HI

Several key frameworks illuminate the evolution of human-AI (HI) collaboration. These frameworks shift the focus from theoretical concepts to actionable design paradigms, emphasizing the creation of systems engineered for deep, symbiotic partnerships with humans.

The Hybrid Intelligence Technology Acceptance Model (HI-TAM)

This model addresses the psychological and practical barriers to HI system adoption, particularly in creative fields where professionals may be apprehensive about AI. HI-TAM posits that successful adoption hinges on two critical elements:

- **Programmable Common Language:** The development of a clear communication channel enabling human experts to convey their goals and constraints to the AI.
- **Human-Centered Continual Training Loop:** The seamless integration of AI training into the expert's natural workflow.

By framing AI as a customizable personal assistant, this approach cultivates a sense of partnership and psychological safety, thereby increasing the expert's willingness to engage with and improve the system (Rafner et al., 2024).

Work System Theory Perspective

This socio-technical approach examines the HI system implementation from an organizational viewpoint. It distinguishes between the roles of the "technical implementer" (AI developer) and the "organizational implementer" (integrator into workflows). This perspective maps challenges and best practices throughout the system's lifecycle, from design to execution and maintenance (Mayer et al., 2024). It underscores that successful HI is not solely a technical endeavor but also an organizational challenge, necessitating careful consideration of workflows, roles, and change management (Mayer et al., 2024).

Thg Homoggnization Risk: A Countgrpoint to thg Crgativg Promisg ofi HI

While promising for enhancing creativity, the paradigm of human-AI (HI) collaboration carries a significant, often overlooked risk: cognitive and cultural homogenization (Agarwal et al., 2024; Rettberg, 2024). A concerning paradox emerges from recent research, indicating that the collective interaction of millions of users with the same foundational generative AI models can result in a convergence of ideas, styles, and solutions (Agarwal et al., 2024; Kumar et al., 2024). Users often face a dilemma: expend effort to articulate unique preferences or readily accept a "good enough" AI-generated output. Rationally, they may gravitate towards the latter (Noy & Zhang, 2023). This can create a "homogenization death spiral": AI-generated content floods the internet, and becomes the training data for the next generation of AI, which in turn produces even more homogenized output, reinforcing mainstream perspectives and marginalizing unique ones (Rettberg, 2024).

A Cross-Domain Analysis ofi Human-AI Collaboration in Practicg

Our framework reveals a powerful explanatory pattern: a precise "paradigm-domain fit" where a domain's primary value driver—efficiency, accountability, or creativity—strongly predicts the dominant human-AI interaction model. An analysis across various sectors demonstrates that the choice of interaction paradigm is not arbitrary but strongly correlated with a given domain's core objectives and constraints. This suggests that as AI becomes more deeply integrated into the economy, we will see an increasing specialization of these interaction paradigms based on industry-specific requirements. This observed 'paradigm- domain fit' provides a strong basis for empirical validation, suggesting that the framework can be used for explanation and prediction, a direction we explore further in our research agenda.

Thg Insufifiicigncy ofi thg "Loop" as an Ethical Safigguard: Collaboration Thgatrg

The notion that human oversight inherently confers ethical legitimacy upon an AI system is a dangerously simplistic assumption. In practice, the "human in the

loop" is often a single individual—a clinician, a content moderator, a manager—who may have a limited understanding of the AI's inner workings, insufficient time to investigate its outputs thoroughly, and their own set of cognitive biases (Griffen & Owens, 2024). This approach can lead to a "scapegoat-in-the-loop" scenario, where the human operator is held responsible for system failures, while the underlying systemic issues in the AI model, its training data, or its organizational deployment go unaddressed (Salloch & Eriksen, 2024).

Furthermore, recent experimental findings suggest that in some cases, the "human-in-the- loop" does not improve decision accuracy but primarily serves to **increase the user's willingness to accept the algorithm's recommendation** (Sele & Chugunova, 2023). One study found that while adding a human monitor increased the uptake of an automated decision support system, it actually *decreased* the quality of the final decisions because humans struggled to appropriately adjust the AI's recommendations, especially the most inaccurate ones (Sele & Chugunova, 2023). This points to the possibility of **"collaboration theatre"**: a performance of oversight that provides psychological comfort and a veneer of legitimacy, making users and organizations more likely to adopt and trust an automated system, even at the cost of accuracy (Rosso, 2024). This "illusion of collaboration" suggests that some systems are designed not for better performance, but for better adoption, using the human's presence as a tool for legitimation rather than for genuine oversight (Jang, 2024).

A Maturity Modgl ofi Human Involvgmgnt in AI Systgms

The technological, interactive, and ethical evolutions discussed throughout this research can be synthesized into a single, cohesive maturity model. This model, visualized in Table 2, provides a narrative arc for the field, offering a clear framework for practitioners and researchers to assess the sophistication of current systems and to guide the design of future

human-AI collaborations. It serves as a diagnostic tool and a conceptual map, linking the paradigms to their corresponding goals, ethical challenges, and mitigation strategies.

Table 2: A Maturity Model of Human Involvement in AI Systems

Stage of Maturity	Primary Paradigm	The goal of Human Involvemen t	Key Ethical Challenge	Ethical Mitigation Strategy
Stage 1: Automation	Human- out- of-	None (Full Automation	"Black Box" opacity;	Post-hoc audits; Impact assessments

	the-loop	)	Algorithmic bias	
Stage 2: Supervision	Human-in-the- Loop (HIL)	Improve AI model performance and accuracy	Biased human input corrupts the model (Garbage-in, Garbage-out)	Data quality control; Annotator diversity; Structured ethical annotation (Savage, 2023; Wang et al., 2023)
Stage 3: Augmentation	AI-in-the-Loop (AI2L)	Enhance human expert decision-making and efficiency	Human over-reliance and automation bias	Explainable AI (XAI); Human-centric evaluation of the socio- technical system; Interface design for critical thinking (Arambepola & Munasinghe, 2021; Salloch & Eriksen, 2024)
Stage 4: Partnership	Hybrid Intelligence (HI)	Co-create novel solutions and achieve synergistic outcomes	Ambiguity of responsibility and accountability for co-created outputs	Clear protocols for shared agency; Dynamic task allocation frameworks
Stage 5: Governance	Participatory System	Ensure long- term alignment with human values, fairness, and safety	Systemic power imbalances; Lack of stakeholder representation	Formal, multi-stakeholder governance bodies; Procedural justice mechanisms (Griffen & Owens, 2024)

.

Discussion: A Rgsgarch Aggnda fior thg Futurg ofi Human-AI Collaboration A Structurgd Rgsgarch Aggnda

This research agenda proposes a framework for understanding the evolving paradigms of human-AI collaboration. Drawing on established theories from teamwork and organizational science, it outlines a lifecycle for human-AI teams, structured around four core concepts: (1) Team Formulation and Coordination, (2) Team Maintenance and Training, (3) Empirical Validation and Evaluation, and (4) Governance and Organizational Integration (Lou et al., 2025). This structure aims to ground future work in existing theoretical understanding.

Tgam Formulation and Coordination

This area concerns the initial design and structuring of the human-AI team, including role allocation and communication protocols.

- **Optimizing Role Specification (HIL, AI2L, HI):** Research is needed on frameworks for allocating roles and responsibilities based on the chosen paradigm. For HIL, this involves optimizing the "supervisor" role to mitigate bias and improve feedback quality. For AI2L, it means designing the "intelligent assistant" to avoid cognitive overload and present information actionably. For HI, this extends to managing dynamic task allocation and shared agency.
- **Developing a "Common Language" for Co-Construction (HI):** How can we develop effective and intuitive "common languages" that allow human experts to communicate complex, domain-specific goals to an AI partner? Research should explore grammar- based methods, visual interfaces, and other modalities for diverse creative and technical domains (Rafner et al., 2024).

Tgam Maintgnancg and Training

This area focuses on the ongoing processes that sustain effective collaboration, including trust-building, learning, and adaptation.

- **Managing Trust and Mitigating Automation Bias (AI2L, HI):** Further research is needed to understand the psychological factors that lead to automation bias and to develop interventions—such as confidence scores, explanations of uncertainty, and interactive training modules—that encourage appropriate reliance on AI suggestions and promote critical thinking (Kumar et al., 2024). For HI systems, this includes fostering psychological safety to encourage adoption by experts who may fear replacement.
- **Optimizing Mutual Learning and Feedback (HIL, HI):** Research is needed on more effective methods for active learning, where the system intelligently queries the human for the most informative labels (Kumar et al., 2024). This includes exploring which types of feedback (e.g., personal-centric vs. community-centric) are most effective at maintaining user engagement and improving model performance. For HI, this extends to studying the bidirectional "hybrid collaborative learning cycle" where both humans and

AI adapt over time (Wiethof & Bittner, 2022).

Empirical Validation and Human-Cgntric Evaluation

A critical direction for future work is the empirical validation of this research framework and the development of robust evaluation methodologies.

- **Developing Human-Centric Evaluation Methods (AI2L, HI):** There is a critical need to develop and validate robust, human-centric methodologies for evaluating the efficacy of the entire socio-technical system, not just the standalone AI model. This requires moving beyond lab-based studies with proxy users to real-world deployments with domain experts and measuring outcomes like decision quality, user trust, and overall system performance (Arambepola & Munasinghe, 2021; Natarajan et al., 2024).
- **Conducting Comparative Case Studies:** Future research should conduct in-depth, comparative case studies to empirically test the framework's "paradigm-domain fit" hypothesis. For example, a study could compare two healthcare systems—one using an AI2L diagnostic tool and another using a more HIL-oriented system—measuring differences in clinical outcomes, workflow efficiency, and practitioner satisfaction. This would provide crucial empirical evidence for the framework's predictive power and practical utility.

Govgrnancg and Organizational Intggration

This area addresses the ethical and structural challenges of embedding human-AI systems into real-world contexts.

- **Designing Scalable Governance Models:** What are practical, scalable, and adaptable models for implementing multi-stakeholder governance in different organizational contexts, from large healthcare systems to technology companies? This includes developing charters, protocols, and decision-making procedures that ensure fairness and accountability (Griffen & Owens, 2024).
- **Facilitating Productive Dialogue:** How can we design processes and tools that effectively manage the power imbalances and epistemic asymmetries between diverse stakeholders? Research is needed on facilitation techniques and collaborative platforms that empower non-experts to contribute meaningfully alongside technical experts (Griffen & Owens, 2024).
- **The Economics and Ethics of Crowd Work (HIL):** Investigating the socio-economic impact of HIL systems on the global workforce of data labelers, with a focus on fair wages, working conditions, and the potential for de-skilling or, conversely, opportunities for new forms of skilled labor (Amershi et al., 2014).

Futurg Controvgrsigs: Bgyond thg Currgnt Aggnda

The research agenda must also anticipate the next wave of fundamental challenges as the field matures. The following controversies are likely to define the future of human-AI collaboration:

- **Cognitive Atrophy vs. Cognitive**

Enhancement: A key irony of automation is that mechanizing routine tasks may deprive users of opportunities to practice their judgment, potentially leading to the deterioration of cognitive faculties (Al-Sibai, 2025). Will deep partnership with AI in the HI paradigm lead to a widespread atrophying of core human cognitive skills, such as critical thinking, as we increasingly offload mental effort (Gerlich, 2025; SFI Health, 2024).

- **The AI Identification Problem:** As AI-generated content becomes indistinguishable from human-created work, reliable identification becomes critical for accountability and trust. However, current AI detection tools are notoriously unreliable, often producing false positives that disproportionately affect non-native English speakers and exhibiting high error rates that render them unusable for high-stakes applications like academic integrity (AI Detectors, 2024; I Digital Strategies, 2024). This raises a crucial governance question: How can we develop a standardized, reliable system to label and categorize distinct AI instances, akin to a software license or serial number, to facilitate identification and ethical oversight? (Gao et al., 2024).
- **The Agency-Performance Trade-Off:** In many high-stakes domains, professionals strongly desire to retain final human agency over decisions (Mayer & Karny, 2025). Yet, studies increasingly show that this does not always lead to the best performance, and in some cases, human intervention can decrease accuracy (Mayer & Karny, 2025; Sele & Chugunova, 2023). How will we navigate this fundamental conflict between the human desire for control and the empirical reality of system performance, especially when lives and livelihoods are at stake? (Watkins, 2025).

Conclusion: From Partnership to Stewardship—Navigating the Cognitive Reshaping of the Hybrid Intelligence Era

The conceptualization of human-AI collaboration has matured significantly from its origins in the simple "Human-in-the-Loop" model. The monolithic application of this term has obscured a rich and diverse reality of interaction, creating ambiguity that hinders both scientific progress and responsible implementation. This research has sought to resolve this ambiguity by proposing a more granular tripartite framework—derived from a systematic literature review—that distinguishes between AI-led supervision (HIL), human-led augmentation (AI2L), and symbiotic partnership (Hybrid Intelligence). This framework, grounded in the locus of control and the primary goal of the interaction, provides a powerful lens for analyzing existing systems and designing future ones.

The future of practical and ethical AI lies not in the relentless pursuit of perfect, human- independent autonomous agents but in the thoughtful and deliberate design of more innovative, more collaborative, and more accountable partnerships between humans and machines (Mayer et al., 2024; Rafner et al., 2024). The research

agenda outlined here underscores that the most pressing challenges are no longer purely algorithmic but are deeply intertwined with human-computer interaction, organizational science, and socio- technical ethics (Kirsten et al., 2025; Natarajan et al., 2024). By embracing this interdisciplinary reality, utilizing more precise conceptual frameworks, and confronting emerging controversies head-on, the research community can better guide the development of AI systems that truly augment human potential and align with our most cherished societal values.

REFERENCES

- Agarwal, D., Naous, T., & Vashistha, A. (2024). *AI suggestions homogenize writing toward Western styles and diminish cultural nuances*. arXiv. https://doi.org/10.48550/arXiv.2409.11360
- AI Detectors: An Ethical Minefield. (2024, December 12). *NIU Center for Innovative Teaching and Learning*. https://citl.news.niu.edu/2024/12/12/ai-detectors-an-ethical- minefield/
- Al-Sibai, N. (2025, February 11). The study finds that people who entrust tasks to AI are losing their critical thinking skills. *Futurism*. https://futurism.com/study-ai-critical-thinking
- Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *AI Magazine, 35*(4), 105–120. https://doi.org/10.1609/aimag.v35i4.2513
- Arambepola, N., & Munasinghe, L. (2021). Human in the loop design for intelligent interactive systems: A systematic review. In *Proceedings of the International Conference on Applied and Pure Sciences (ICAPS 2021-Kelaniya)* (Vol. 1, p. 225). Faculty of Science, University of Kelaniya, Sri Lanka. http://repository.kln.ac.lk/handle/123456789/24082
- Bastani, H., Bayati, M., & Khosravi, K. (2017). *Mostly exploration-free algorithms for contextual bandits*. arXiv. https://doi.org/10.48550/arXiv.1704.09011
- Chen, X., Wang, X., & Qu, Y. (2023). Constructing ethical AI based on the "human-in- the-loop" system. *Systems, 11*(11), 548. https://doi.org/10.3390/systems11110548
- Dutta, M., Murray, L. L., & Stark, B. C. (2024). The relationship between executive functioning and narrative language abilities in aphasia. *American Journal of Speech- Language Pathology, 33*(5), 2500–2523. https://doi.org/10.1044/2024_AJSLP-23-00314
- Falk, J., Chen, Y., Rafner, J., Zhang, M., Bjerva, J., & Nolte, A. (2025). *How do hackathons foster creativity? Towards AI collaborative evaluation of creativity at scale*. arXiv. https://doi.org/10.48550/arXiv.2503.04290
- Gao, D. K., Haverly, A., Mittal, S., Wu, J., & Chen, J. (2024). AI ethics: A bibliometric analysis, critical issues, and key gaps. *International Journal of Business Analytics, 11*(1), 1-19. https://doi.org/10.4018/IJBAN.338367

- Gerlich, A. (2025). The impact of AI tools on critical thinking and cognitive offloading: A cross-sectional study. *Social Sciences & Humanities Open, 15*(1), 6.
- Griffen, Z., & Owens, K. (2024). From "human in the loop" to a participatory system of governance for AI in healthcare. *The American Journal of Bioethics, 24*(9), 81–83. https://doi.org/10.1080/15265161.2024.2377119
- iDigitalStrategies. (2024). *Unraveling the quandary: The problem with AI-generated content detectors*. https://www.idigitalstrategies.com/blog/problem-with-ai- generated-content-detectors/
- Jang, J. (2024, June 10). *Some thoughts on human-AI relationships*. OpenAI. https://community.openai.com/t/some-thoughts-on-human-ai-relationships/1279464
- Jayapradha, J., Sujin, B. B., Rani, M. J., Lotus, R., & Ahamed, A. F. (2024). Human-AI collaboration via a hybrid intelligent system for safe autonomous driving. *Nanotechnology Perceptions, 20*(S7), 133–147.
- Kang, H., Qian, X., Hope, T., Shahaf, D., Kittur, A., & Chan, J. (2022). Augmenting scientific creativity with an analogical search engine. *ACM Transactions on Computer- Human Interaction, 29*(6), 1–41. https://doi.org/10.1145/3530013
- Kehinde-Awoyele, A. A., Adeowu, W. A., & Oladejo, B. (2024). Enhancing classroom learning: The impact of AI-based instructional strategies on student engagement and outcomes. *International Journal of Research and Innovation in Social Science, 8*(12), 5732-5742.
- Kirsten, L., Lou, B., Lu, T., Raghu, T. S., & Zhang, Y. (2025). *Unraveling human-AI teaming: A review and outlook*. arXiv. https://arxiv.org/abs/2504.05755
- Kumar, A., Zhang, N., & Zhu, T. (2024). Enhancing AI reliability in public health with human-in-the-loop approaches. *American Journal of Public Health, 114*(S6), S476– S479. https://doi.org/10.2105/AJPH.2024.307888
- Kumar, V., Talwar, S., & Doshi, P. (2024). *The diversity-innovation paradox in generative AI*. arXiv. https://doi.org/10.48550/arXiv.2405.13868
- Lou, B., Lu, T., Raghu, T. S., & Zhang, Y. (2025). *Unraveling human-AI teaming: A review and outlook*. arXiv. https://arxiv.org/abs/2504.05755
- Malone, T. W., Almaatouq, A., & Vaccaro, M. (2025, February 3). When humans and AI work best together — and when each is better alone. *MIT Sloan Management Review*.
- Mayer, J. F., Madden, E. B., Mozeiko, J., Murray, L. L., Patterson, J. P., Purdy, M., Sandberg, C. W., & Wallace, S. E. (2024). Generalization in aphasia treatment: A tutorial for speech-language pathologists. *American Journal of Speech-Language Pathology, 33*(1), 57–73. https://doi.org/10.1044/2023_AJSLP-23-00192
- Mayer, L. W., & Karny, S. (2025). *Human-AI collaboration: Trade-offs between performance and preferences*. arXiv. https://doi.org/10.48550/arXiv.2503.00248
- Medepalli, S. (2025). Human-AI

collaboration (HAIC): The rise of hybrid intelligence in modern software development. *Journal of International Research for Engineering & Management, 10*(1). https://doi.org/10.5281/zenodo.14743406

- Natarajan, S., Mathur, S., Sidheekh, S., Stammer, W., & Kersting, K. (2024). *Human-in- the-loop or AI-in-the-loop? Automate or collaborate?* arXiv. https://doi.org/10.48550/arXiv.2412.14232
- Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187-192.
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, n71. https://doi.org/10.1136/bmj.n71
- Rafner, J., Sherson, J., & Qualter, P. (2024). Creativity in the age of generative AI. *Current Directions in Psychological Science, 33*(2), 108–116. https://doi.org/10.1177/09637214231222549
- Rettberg, J. W. (2024). To counter AI's cultural biases, we need to teach it to tell new stories. *Issues in Science and Technology*. https://issues.org/generative-ai-cultural- narratives-rettberg/
- Rosso, C. (2024, November 11). How synergistic is the combo of AI and humans? *Psychology Today*. https://www.psychologytoday.com/us/blog/the-future- brain/202411/how-synergistic-is-the-combo-of-ai-and-humans
- Salloch, S., & Eriksen, A. (2024). What are humans doing in the loop? Co-reasoning and practical judgment when using machine learning-driven decision aids. *The American Journal of Bioethics, 24*(9), 67–78. https://doi.org/10.1080/15265161.2024.2353800
- Savage, T. (2023). Human-in-the-loop problem-solving with artificial intelligence. *Academy of Management Review*. https://doi.org/10.5465/amr.2021.0421
- Savage, T., & del Rio Chanona, E. A. (2023). *Expert-guided Bayesian optimization for human-in-the-loop experimental design of known systems*. arXiv. https://doi.org/10.48550/arXiv.2312.028552
- Sele, D., & Chugunova, M. (2023). *Putting a human in the loop: Increasing uptake, but decreasing accuracy of automated decision-making* (Discussion Paper No. 438). Collaborative Research Center Transregio 190, Ludwig-Maximilians-Universität München and Humboldt-Universität zu Berlin.
- SFI Health. (2024). *The impact of AI on cognitive function: Are our brains at stake?* https://www.sfihealth.com/news/the-impact-of-ai-on-cognitive-function-are-our- brains-at-stake
- Singhal, M., Ling, C., Paudel, P., Thota, P., Kumarswamy, N., Stringhini, G., & Nilizadeh, S. (2023). SoK: Content moderation in social media, from

guidelines to enforcement and research to practice. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)* (pp. 488-506). IEEE. https://doi.org/10.1109/eurosp57164.2023.00056

- Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology, 8*(1), 45. https://doi.org/10.1186/1471-2288-8-45
- Wang, X., Chen, X., & Qu, Y. (2023). Constructing ethical AI based on the "human-in- the-loop" system. *Systems, 11*(11), 548. https://doi.org/10.3390/systems11110548
- Watkins, E. A. (2025). *How to resolve the five trade-offs of AI*. IMD. https://www.imd.org/ibyimd/brain-circuits/how-to-resolve-the-five-trade-offs-of-ai/
- Wiethof, C., & Bittner, E. A. C. (2022). Toward a hybrid intelligence system in customer service: collaborative learning of human and AI. *Proceedings of the 55th Hawaii International Conference on System Sciences*.