Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 6 Issue 1 (2026)

# Beyond the Loop: A Tripartite typology of human-machine collaboration.

**Sagar Patil[1],Suneel Sharma[2], Mehregan Mahdavi[3]**

*[1]DBA18, SP Jain School of Global Management.*
*[2](Principal Supervisor).*
*[3](Co Supervisor).*

## Abstract
The “human in the loop” (HIL) label that currently exists has become an exaggerated umbrella over critical distinctions within human–AI partnerships. This will provide a tripartite schema that categorises the HIL (AI led, automation first), AI in the loop (AI2L; human led, augmentation first) and Hybrid Intelligence/Collaborative Intelligence (HI/CI; co creative partnership). We suggest a paradigm–domain correspondence: efficiency driven contexts favor HIL, accountability driven domains favor AI2L, and creativity driven domains favor HI/CI. We integrate evidence to support a continuing agency–performance tradeoff in high stakes environments, the AI identification challenge, and the homogenization risk of scaled generative models on culture. We suggest for move from a brittle “human in the loop” backstop to participatory governance — organized, multi-stakeholder overview of the entire lifecycle. A four-part research agenda and a maturity model of human involvement are provided to inform empirical validation and ethical deployment. This analysis lays the groundwork for more sophisticated analyses and calls for interdisciplinary, field based evaluation of human–AI systems to ensure they augment human flourishing in ways consistent with the principle of preserving diversity of thought and autonomy of judgement.

## Introduction
"In the loop" used to describe the role of humans in AI systems; today is expanded to include annotation, decision support and co creative work, to the extent that design intent and accountability mix. Such a conceptual vagueness is part of this performance paradox that many researchers have identified – namely that human–AI teams rarely outperform AI, and that human involvement can sometimes degrade accuracy (Sele & Chugunova, 2023). Professionals are incentivized in life changing environments to keep final control, which runs counter to the best available evidence (Mayer & Karny, 2025; Watkins, 2025). Taken together, these tensions represent a deeper conceptual crisis and the need for more accurate typologies and ethical guardrails.

## Thesis and Contribution
Focused on authority allocation, goals, and the interaction dynamic between participant and agent, this study proposes a three tier framework of human-AI collaboration--HIL, AI2L, and HI/CI--that can support such collaboration. It provides paradigm-domain correspondence, critiques the ethical insufficiency of "a human in the loop" as a shield, and proposes participatory governance as a systemic alternative. Filled with interdisciplinary literature, the framework offers a tool for analysis that gives guidance on how to design research designs and systems, as well as policy directions. This framework is relevant now that AI is coming onto the scene in areas like governance, social welfare or health care. This background gives us a clear idea of the conceptual landscape.

## Background and Conceptual Landscape

Earlier AI workflows put humans as supervisors, teachers, or oracles, mainly involved in data curation, labeling, and corrective feedback (Amershi et al., 2014; Bastani et al., 2017). With AI moving into high-stakes fields, human experts began relying on AI's decision support without relinquishing responsibility—an AI2L outlook common in medicine and finance (Arambepola & Munasinghe, 2021; Wang et al., 2023). The rise of generative models allowed for iterative, co creative workflows, wherein humans and machines collaboratively ideate and refine outputs (Kang et al., 2022; Rafner et al., 2024; Medepalli, 2025; Wiethof & Bittner, 2022); a defining feature of HI/CI.

But this evolution highlighted inconsistencies among goals, roles and evaluation. In some cases, adding humans increases uptake while decreasing accuracy (Sele & Chugunova, 2023). For others, it is easier for users to embrace "good enough" AI outputs and not to invest effort to articulate their individual preferences; potentially resulting in greater convergence around style and content (Noy & Zhang, 2023). These are some of what we can use to promote distinctions that link locus of control and roles to proper performance and ethical metrics.

## A Tripartite Typology of Human–AI Collaboration

Human in the Loop (HIL: AI led, automation first).

- Locus of control: The AI system is the primary actor; humans intervene at defined points (e.g., labeling, auditing, corrective feedback).
- Purpose: Improve model accuracy, efficiency, and scalability.
- Evaluation: AI centric metrics (accuracy, precision/recall, F1).
- Risks: Data quality, annotator bias; scalability and fairness in crowd work.
- Representative contexts: Large scale content moderation; autonomous data processing (Singhal et al., 2023).

AI in the Loop (AI2L: human led, augmentation first).

- Locus of control: Human experts lead; AI provides analyses, recommendations, and uncertainty indicators.
- Purpose: Improve decision quality and efficiency while retaining human accountability.
- Evaluation: Human and workflow centric metrics (decision quality, user trust, cognitive load, disparities).
- Risks: Automation bias, cognitive overload, "collaboration theatre" where nominal oversight legitimizes systems without improving outcomes (Salloch & Eriksen, 2024).
- Representative contexts: Healthcare, finance, and other accountability driven domains (Griffen & Owens, 2024).

Hybrid/Collaborative Intelligence (HI/CI: co creative partnership).

- Locus of control: Shared; dynamic task allocation and iterative co construction.
- Purpose: Novel ideation and synergistic outcomes beyond either partner alone.
- Evaluation: Creativity, originality, user satisfaction, and real world impact at the team level.
- Risks: Diffuse accountability; cultural and cognitive homogenization as outputs converge (Agarwal et al., 2024; Rettberg, 2024; Kumar, Talwar, & Doshi, 2024).
- Representative contexts: Generative design, software development, scientific ideation, creative industries (Kang et al., 2022; Medepalli, 2025; Rafner et al., 2024).

Paradigm–Domain Correspondence

We hypothesize a systematic mapping between domain imperatives and collaboration paradigms:

- Efficiency focused domains favor HIL. Where throughput and scale dominate (e.g., social media moderation, autonomous pipelines), human effort is channeled to lift model performance and keep costs manageable (Singhal et al., 2023).
- Accountability focused domains favor AI2L. In healthcare and finance, decision rights and liability norms anchor humans as final authorities, with AI supporting situational awareness and analysis (Griffen & Owens, 2024; Arambepola & Munasinghe, 2021).
- Creativity oriented domains favor HI/CI. When novelty is the core value (e.g., generative design, software engineering, research ideation), co construction and flexible handoffs are more appropriate (Kang et al., 2022; Medepalli, 2025; Wiethof & Bittner, 2022).

This correspondence both clarifies design choices and predicts adoption: efficiency tracks to HIL pipelines and model metrics; accountability to human factors evaluation and documentation; creativity to shared agency tooling and team level assessments.

**The Agency–Performance Trade-off**

In many high stakes contexts, professionals strongly prefer to retain final control (Mayer & Karny, 2025). Yet studies show this preference does not always yield best performance; in some cases, human intervention decreases accuracy relative to AI alone (Sele & Chugunova, 2023). This creates a fundamental tension between legal/ethical agency and empirical performance (Watkins, 2025). Design and governance responses differ by paradigm:

- HIL: Strengthen annotation protocols, quality controls, and annotator diversity; ensure fair labor practices.
- AI2L: Calibrate reliance with uncertainty communication, interpretable recommendations, and training that fosters critical engagement and mitigates automation bias (Salloch & Eriksen, 2024).
- HI/CI: Establish norms for shared agency, traceability, and recourse; design handoff mechanisms that make joint authorship and accountability auditable.

**The AI Identification Problem**

As AI generated content becomes harder to distinguish from human work, some institutions look to detectors. Evidence shows these tools are unreliable, have high error rates, and disproportionately flag non native English writers—making them unsuitable for high stakes uses such as academic integrity (AI Detectors, 2024; iDigitalStrategies, 2024). A more promising approach is standardized provenance and process transparency: system level labels or signatures to identify model families/versions and documented workflows indicating when and how AI contributed, balancing transparency with privacy and practicality (Gao et al., 2024).

**Cultural and Cognitive Homogenization: A Counterpoint to Creative Promise**

Generative AI's creative boost can be accompanied by convergence risks. Widely used, Western centric models can nudge users toward Western idioms and references, diluting local nuance (Agarwal et al., 2024; Rettberg, 2024). Faced with effort–benefit trade offs, users may accept "good enough" outputs, amplifying stylistic and conceptual homogenization (Noy & Zhang, 2023; Kumar, Talwar, & Doshi, 2024). Over time, heavy reliance may also reshape cognition via offloading, warranting longitudinal study and mitigations (Gerlich, 2025; SFI Health, 2024; Al Sibai, 2025). Mitigation strategies include:

- Data and model pluralism to broaden perspectives.

- Interfaces that nudge diversity by surfacing counter styles and minority viewpoints.
- Evaluation that values novelty and diversity, not just productivity.
- Participatory oversight with cultural experts and affected communities to shape curation and assessment.

**From "Human in the Loop" to Participatory Governance**

Invoking "a human in the loop" often serves as a thin ethical assurance. In practice, it can devolve into collaboration theatre or participation washing that adds a veneer of legitimacy without redistributing power or ensuring accountability (Salloch & Eriksen, 2024; Griffen & Owens, 2024). We argue for participatory governance: multi stakeholder, lifecycle oversight with structured accountability and enforcement capacity. Key elements include:

- Clear decision rights, escalation paths, and audit trails aligned to HIL, AI2L, and HI/CI workflows.
- Lifecycle reviews: impact assessments, post deployment monitoring, and red team exercises. • Inclusion of end users and affected communities in requirements, evaluation, and recourse.
- System level socio technical metrics (e.g., disparities, trust, decision quality), not only model scores (Gao et al., 2024; Wang et al., 2023).

**Methods and Scope**

This paper synthesizes interdisciplinary literature across AI, HCI, organizational studies, and ethics, drawing on systematic review practices and thematic synthesis to articulate a testable framework (Page et al., 2021; Thomas & Harden, 2008). The aim is theory building rather than a meta analysis; the research agenda below specifies empirical strategies to validate and extend the framework.

**A Structured Research Agenda.**

1. Team creation and coordination:

Codify roles and handoffs by paradigm: HIL teacher/oracle protocols for data quality; AI2L assistants that surface actionable insights without overload; HI/CI workflows for shared agency and dynamic assignment of tasks (Arambepola & Munasinghe, 2021; Natarajan et al., 2024).
Develop co-construction languages and multimodal interfaces for HI/CI (Kang et al., 2022; Rafner et al., 2024).

2. Team maintenance and training:

Challenge trust and automation bias by giving a sense of uncertainty, justifying and applying it in an interpretable manner and employing scenario based training that maintains critical thinking (Salloch & Eriksen, 2024; Kumar, Zhang, & Zhu, 2024). Create a sense of psychological safety so that people with expertise can be taught to embrace it without fear of displacement.

3. Validation and experimentation:

Move from lab analogues to the field using an authentic subject matter expert; to track decisions, cognitive load, work quality, workflow, and fairness influence (Lou et al., 2025; Kirsten et al., 2025).

Contrast case studies to benchmark paradigm–domain fit (e.g., testing AI2L vs. HIL against each other in similar clinical tasks).

4. Organizational integration and governance:
Establish functional participatory governance boards, recourse procedures, and ongoing audits with enforcement (Griffen & Owens, 2024).
Set fair conditions for HIL annotators and clarify liability and record keeping across teams for AI2L and HI/CI.

**A Model for Human Involvement Maturity**
We integrate technological, interactive and ethical development into a maturity model from which to assess and redesign:
- Level 1: Total Autonomy (no human involvement included).
  - Objective: Scale/swift; Risks: Opacity, embedded bias; Mitigation: auditing, impact calculations.
- Level 2: Supervised Automation (HIL, AI governed.)
  - Objective: Increase model efficiency through human feedback; Risks: biased/inconsistent inputs; Mitigation: data quality controls, annotator diversity, structured guidelines (Savage, 2023; Wang et al., 2023).
- Level 3: Augmentation (AI2L; human-managed).
  - Objective: Improve human decisions; Risks: automation bias, cognitive overload; Mitigation: explainability, uncertainty communication, human centered interface evaluations (Arambepola & Munasinghe, 2021; Salloch & Eriksen, 2024).
- Level 4: Collaboration — HI/CI.
  - Objective: Work together toward new solutions; Risks: murky obligations; Mitigation: procedures for shared responsibility, openness in handoffs, auditable ownership (Rafner et al., 2024).
- Level 5: Participatory Governance
  - Goal: Institutionalize fairness and accountability; Risks: formalism without redistribution of power; Mitigation: inclusiveness; continuous oversight with measurable socio technical results; (Griffen & Owens, 2024; Gao et al., 2024)

**Domain Illustrations.**
- Scalable content/data processing (HIL): The moderation of content needs annotation quality, consistency and throughput; human observers help in keeping models aligned as norms evolve (Singhal et al., 2023).
- Healthcare/finance (AI2L): Human clinicians or analysts are at the forefront; AI augments judgment via risk scores, differential diagnoses, or anomaly detection; interfaces need to calibrate reliance and document decisions (Griffen & Owens, 2024).
- Generative design and software engineering (HI/CI): Joint creative workflows facilitate exploration and code suggestions, analogical search; on-the-fly handoff and provenance support not just innovation but also accountability (Kang et al., 2022; Medepalli, 2025; Wiethof & Bittner, 2022).

**Limitations and Future Research Directions**
This conceptual framework is a conceptual basis and needs to achieve empirical confirmation across tasks, domains and cultures. Priorities include:

- Longitudinal research on cognitive offloading and skill development with and without AI in real time (Gerlich, 2025; SFI Health, 2024; Al Sibai, 2025).
- Cross cultural comparative analysis of homogenization risks and mitigation strategies (Agarwal et al., 2024; Rettberg, 2024).
- Field experiments examining other provenance alternatives to substitute unreliable detectors (AI Detectors, 2024; iDigitalStrategies, 2024; Gao et al., 2024).
- Case studies comparing paradigms vs domain fit and agency v performance cost in real workflow (Kirsten et al., 2025; Lou et al., 2025).

## Conclusions

The field is at a crossroads where the clarity of ideas can both advance science and make AI more responsible. Reaching out beyond a monolithic HIL designation, the future of HIL, AI2L, and HI/CI must reflect the roles, assessment, and ethical protections relevant to domain objectives. Conclusiveness should supplant thin guarantees of 'a human in the loop'; governance should be participatory and has oversight throughout the loop. By utilizing these interfacilitative strategies, and through attention to field based validation, we can design human-AI systems that are truly designed for human flourishing, but without robbing humans of the multiplicity of ideas and autonomous judgment that makes humans tick.

## References

1. 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
2. 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/
3. Al-Sibai, N. (2025, February 11). Study finds that people who entrust tasks to AI are losing their critical thinking skills. Futurism. https://futurism.com/study-ai-critical-thinking
4. 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
5. Arambepola, N., & Munasinghe, L. (2021). Human in the loop design for intelligent interactive systems: A systematic review. In Proceedings of ICAPS 2021—Kelaniya (Vol. 1, p. 225). University of Kelaniya. http://repository.kln.ac.lk/handle/123456789/24082
6. Bastani, H., Bayati, M., & Khosravi, K. (2017). Mostly exploration-free algorithms for contextual bandits. arXiv. https://doi.org/10.48550/arXiv.1704.09011
7. 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
8. Gerlich, A. (2025). The impact of AI tools on critical thinking and cognitive offloading: A cross-sectional study. Social Sciences & Humanities Open, 15(1), Article 6.
9. 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
10. iDigitalStrategies. (2024). Unraveling the quandary: The problem with AI-generated content detectors. https://www.idigitalstrategies.com/blog/problem-with-ai-generated-content-detectors/

11. 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.
12. 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
13. 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.
14. 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
15. 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
16. Kumar, V., Talwar, S., & Doshi, P. (2024). The diversity–innovation paradox in generative AI. arXiv. https://doi.org/10.48550/arXiv.2405.13868
17. 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
18. 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.
19. Mayer, L. W., & Karny, S. (2025). Human–AI collaboration: Trade-offs between performance and preferences. arXiv. https://doi.org/10.48550/arXiv.2503.00248
20. 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
21. 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
22. Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Science, 381(6654), 187–192.
23. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, n71. https://doi.org/10.1136/bmj.n71
24. 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
25. 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/
26. 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
27. Savage, T. (2023). Human-in-the-loop problem-solving with artificial intelligence. Academy of Management Review. https://doi.org/10.5465/amr.2021.0421
28. 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.

29. 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
30. 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
31. 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
32. 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
33. 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/
34. 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.