European Economic Letters
ISSN 2323-5233
Vol 16, Issue 1 (2026)
http://eelet.org.uk

From Loop To Partnership: A Model For a Changing Paradigm Of Human–Ai Partnership.

Sagar Patil,
SP Jain School of Global Management.

Suneel Sharma
(Supervisor).

Mehregan Mahdavi
(Co-Supervisor)

Abstract
Human-in-the-loop (HIL) once was described to clarify the role of human in AI systems, but has been extended to cover interactions that represent divergent forms, compromising our designers intentions and ethical guardrails. This paper seeks to resolve this conceptual ambiguity by proposing a tripartite framework—HIL (AI-led, automation-first), AI-in-the-loop (AI2L; human-led, augmentation-first), and Hybrid Intelligence (HI; co-creative partnership)—and by suggesting a paradigm–domain fit that associates each paradigm with objectives pertaining to efficiency, accountability, or creativity. We synthesize evidence of the performance paradox—human–AI teams often lag behind AI alone—and the agency–performance trade-off that emerges in high-stakes settings. The paper contends that ethical oversight must transform from a "thin human in the loop" defence to participatory governance, which permeates multi-stakeholder accountability, embedding accountability to take effect across the lifecycle. We tackle two significant problems: AI-generated content detectors that are unreliable and cultural homogenization risks as generative models grow. A structured research agenda emphasizes team formation, maintenance, evaluation in the wild and governance. This framework is designed to help both researchers and practitioners build human–AI systems which are compatible with the intended goals of the domain, and ensure that bias, overabundance, and dispersed responsibility are minimized.

Key Words: Human–AI Collaboration, Participatory Governance, Performance Paradox, Generative AI

Introduction
Human to AI cooperation is expanding in various domains, but the generic definition of 'human-in-the-loop' has run the risk of being conceptually overstretched. It embraces annotation workflows, decision support, and co-creative teaming, without even naming responsibility (who is in charge and what the human will do), or oversight mechanisms (over time) (Griffen & Owens, 2024). This ambiguity has real-world implications: the 'performance paradox' argues that human–AI teams often are not superior to AI only (Malone et al., 2025), and sometimes human intervention cuts down on accuracy (Sele & Chugunova, 2023; Mayer & Karny, 2025). We argue that the area requires transparent paradigms that link roles and objectives with assessment and stewardship.

We develop a tripartite model to define AI-led HIL, human-led AI2L, and co-creative HI, and a paradigm–domain fit of efficiency, accountability and creativity for each model. We contend

that moral protections should evolve from a brittle "human-in-the-loop" backstop to participatory governance, structured, multi-stakeholder accountability throughout the lifecycle (Griffen & Owens, 2024). This paper also integrates emerging risks, such as cultural homogenization and untrustworthy AI detection tools (Agarwal et al., 2024; AI Detectors, 2024; iDigitalStrategies, 2024), and sets forth an agenda for future work in the socio-technical domains of collaboration.

Methodology

We performed a systematic review to find relevant studies across AI, HCI, organizational science, and ethics (Page et al., 2021). The researchers synthesized qualitative evidence using thematic synthesis techniques (Thomas & Harden, 2008). Gaps and trends were guided by thematic summaries of AI ethics insights (Gao et al., 2024). This review provides a summary of conceptual and empirical work undergirding our framework and application in areas of efficiency, accountability, and creativity-driven work.

Background and Related Work

Traditionally HIL positions humans as supervising, training or oracles whose work allows AI work to get better through labeling/auditing/feedback loops (Amershi et al., 2014; Bastani et al., 2017). AI2L augments human experts with AI analyses, uncertainty estimates and decision aids while keeping human responsibility of these models, an approach that is prevalent in high stakes environments like healthcare and finance (Arambepola & Munasinghe, 2021). The emergence of generative AI makes it possible to apply HI workflows that enable humans and the AI to co-create pieces of artifacts and iterate on drafts, code, models, or research input (Kang et al., 2022; Medepalli, 2025; Rafner et al., 2024). These findings complicate assumptions about synergy.

Whereas generative AI can help boost individual productivity on many writing tasks (Noy & Zhang, 2023), the general efficacy of human–AI teams is found to be lower than that alone (Malone et al., 2025). In regulated domains, human intervention can hinder accuracy, in particular as the adjustments needed for AI recommendations and management of cognitive load exist and challenge the effectiveness of human-assisted AI interventions (Mayer & Karny, 2025; Sele & Chugunova, 2023). These insights inspire explicit delineation of responsibilities, standards of assessment and ethical scrutiny with regard to paradigms (Natarajan et al., 2024).

A Tripartite Framework

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

- Locus of control: AI is the engine; humans contribute at specific points to increase model performance.
- Role of Human: Supervisor/teacher/oracle (e.g., annotation, auditing, corrective feedback) (Amershi et al., 2014; Bastani et al., 2017).
- Main aim: The accuracy, efficiency and scalability of the model. • Evaluation: AI-centric metrics (accuracy, precision/recall, F1).
- Key risks: Data quality, annotator bias, scalability.
- Example domain: Large-scale content moderation & autonomous data processing pipelines (SoK: Content Moderation in Social Media, 2023).

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

- Locus of control: Human expert leads; AI supports analysis, suggestions, and uncertainty communication (Arambepola & Munasinghe, 2021).
- Human role: Decision maker with legal and ethical responsibility. • Goal: Improve decision quality and efficiency while preserving accountability.
- Evaluation: Human-centric and socio-technical outcomes (decision quality, trust, cognitive load, workflow performance) (Natarajan et al., 2024).
- Risks: Automation bias and cognitive overload; "collaboration theatre" when human presence legitimizes a system without improving accuracy (Rosso, 2024; Salloch & Eriksen, 2024).
- Example domains: Healthcare and finance (Griffen & Owens, 2024; Arambepola & Munasinghe, 2021)

Hybrid Intelligence (HI: co-creative partnership).

- Locus of control: Shared; agency and task allocation shift dynamically across the workflow.
- Human and AI roles: Iterative co-construction, interpretation, and refinement (Rafner et al., 2024).
- Primary goal: Novel ideation and synergistic outcomes neither partner could produce alone. • Evaluation: Creativity, originality, user satisfaction, real-world impact.
- Risks: Diffuse accountability and cultural homogenization as model outputs converge (Agarwal et al., 2024; Rettberg, 2024; Kumar et al., 2024).
- Example domains: Generative design, software development, research ideation, hackathon collaboration (Kang et al., 2022; Medepalli, 2025; Falk et al., 2025).

Paradigm–Domain Fit.

We hypothesize systematic alignment between domain imperatives and collaboration paradigms:

- Efficiency-driven domains (e.g., social media moderation, autonomous data processing) favor HIL. The goal is throughput and scale; humans correct, curate, and supervise to elevate model performance (SoK: Content Moderation in Social Media, 2023).
- Accountability-driven domains (e.g., healthcare, finance) favor AI2L. Human experts retain decisional authority to satisfy legal, professional, and ethical norms; growing scrutiny pushes beyond augmentation to structured, multi-stakeholder oversight (Griffen & Owens, 2024; Arambepola & Munasinghe, 2021).
- Creativity-driven domains (e.g., generative design, modern software workflows, research ideation) favor HI, where iterative co-creation drives value and evaluation must reflect team-level creativity and impact (Kang et al., 2022; Medepalli, 2025; Rafner et al., 2024; Falk et al., 2025).

The Performance Paradox and Agency–Performance Trade-Off

Long-term, sustained evidence shows that human–AI teams often underperform AI alone (Malone et al., 2025). In high-stakes contexts, experts insist on final control, a defensible preference that can reduce accuracy in certain tasks (Mayer & Karny, 2025; Sele & Chugunova, 2023). This tension pits professional agency and legal responsibility against empirical performance. Navigating this trade-off requires paradigm-specific interventions: HIL must

strengthen annotation protocols and bias controls; AI2L must calibrate reliance with uncertainty estimates, explanations, and human-centered interfaces; HI must establish norms and audit trails for shared agency and responsibility (Natarajan et al., 2024; Watkins, 2025).

From "Human in the Loop" to Participatory Governance

While "Human in the loop" is frequently invoked as an ethical safeguard, it can devolve into collaboration theatre or participation washing, where token oversight absorbs blame without structural accountability (Griffen & Owens, 2024; Salloch & Eriksen, 2024). We argue for participatory governance: multi-stakeholder, lifecycle oversight with clear roles for developers, domain experts, end-users, frontline staff, IT/data teams, ethicists, and regulators (Griffen & Owens, 2024). Structured accountability should define decision rights, escalation paths, and audit trails; governance processes should include impact assessments, post-deployment monitoring, and red team exercises; and system-level evaluation must track disparities, trust, and decision quality—not just model metrics (Gao et al., 2024; Wang et al., 2023).

The AI Identification Problem

The problem with the recognition of an AI. With AI-generated works no longer different even on paper from human writing, some schools seek detectors which can have high error rates, false positives and disproportionally high nonnative writer misclassifications (AI Detectors, 2024; iDigitalStrategies, 2024)—a problem especially relevant to high stakes areas such as academic integrity. Standardized provenance and transparent process documentation are the better trajectory: system-level labels or signatures for the family and versions of models, along with auditable workflows that inform when and how the AI contributed (Gao et al., 2024).

Risks of Cultural and Cognitive Homogenization

The potential for generative AI to be a creativity boon is underscored by convergence threats. There has been evidence that widely adopted, Western-centric models serve to push users toward Western idioms and references, reducing local nuance (Agarwal et al., 2024; Rettberg, 2024). In practice, most users settle for "good-enough" outputs and do not attempt to articulate unique user preferences, which further increases homogenization (Noy & Zhang, 2023; Kumar et al., 2024).

Over the years, greater dependence on AI can lead to cognitive reshaping through offloading and thus require longitudinal studies and mitigations to protect critical thinking (Gerlich, 2025; SFI Health, 2024; Al Sibai, 2025). Mitigation strategies might include data and model pluralism, interface nudges to consider stylistic diversity, creativity-focused evaluation, and participatory oversight involving cultural experts and affected communities (Rettberg, 2024).

A Structured Research Agenda

Team building and alignment.

- Specify roles and hand-offs that align with paradigm: HIL teacher/oracle protocols for data quality; AI2L assistants that deliver actionable insights without overwhelming; HI workflows that deliver shared agency and iterative co-creation (Arambepola & Munasinghe, 2021; Natarajan et al., 2024).
- Common languages development (HI) Multimodal interfaces should be created allowing specialized teams to transfer domain goals effortlessly and intuitively through common languages (Rafner et al., 2024).

Training and maintenance for team.

- Tackle trust and automation bias through uncertainty communication, explainability and scenario-based training for appropriate dependence and critical thinking (Kumar et al., 2024; Salloch & Eriksen, 2024).
- Create psychological safety around HI to enable adoption amongst experts who may be nervous about displacement.

Empirical validation and evaluation

- Go from lab proxies to real-world deployments with domain experts measuring the quality of decisions made, cognitive load, trust from users and workflow outcomes on the process (Arambepola & Munasinghe, 2021; Natarajan et al., 2024).
- Conduct comparative case studies to test the paradigm–domain fit hypothesis (e.g. AI2L vs HIL in similar healthcare contexts) (Lou et al., 2025; Kirsten et al., 2025).

Governance and organizational integration.

- Implement participatory governance boards, recourse mechanisms, and continuous audits (Griffen & Owens, 2024).
- For HIL, ensure ethical treatment and fair payments for annotators; for AI2L and HI, define liability, escalation and documentation across teams (Amershi et al., 2014).

A Maturity Model of Human Participation

- Stage 1: Automation (human-out-of-the-loop).
  - Goal: Scale and speed.
  - Risk: Opacity and embedded bias.
  - Mitigation Strategy: Post-hoc audits and impact assessments.
- Stage 2: Supervision (HIL).
  - Objective: Enhance model performance with human input.
  - Risk: Stereotyped or ambiguous human contributions (garbage-in, garbage-out).
  - Remediation: Data quality controls, heterogeneity in annotators, structurally organized ethical annotation (Savage, 2023; Wang et al., 2023).
- Stage 3: Augmentation (AI2L).
  - Objective: Improve the human decision making and the functionality.
  - Risk: Automation bias and cognitive overload.
  - Mitigation: Explainable AI, uncertainty communication, human centric interface evaluation (Arambepola & Munasinghe, 2021; Salloch & Eriksen, 2024).
- Stage 4: Partnership (HI).
  - Goal: Co-create the freshness and successfulness of a combined strategy.
  - Risk: Unclear ownership of co-written products.
  - Mitigation: Explicit measures for distributed agency and distributed tasks with audit trails (Rafner et al., 2024).
- Stage 5: Governance (the participatory system)
  - Goal: Create fairness and accountability throughout the lifetime.
  - Risk: A formalism that lacks redistribution of power or continuous review.
  - Mitigation: Multi-stakeholder monitoring, lifecycle reviews, formalised responsibilities, and socio-technical aspects (Griffen & Owens, 2024; Gao et al., 2024).

Conclusion

The monolithic HIL construct hides significant differences in goals, roles, and risks among human–AI systems. A more nuanced framework — HIL (AI-led), AI2L (human-led), and HI (co-creative)— elucidates evaluation criteria and ethical guardrails and forecasts adoption patterns across efficiency-, accountability-, and creativity-driven domains. As hybrid systems spread, ethical assurances will need to evolve from an abstract construct of "a human in the loop" to participative governance that shares power, sets responsibility and observes socio-technical outcomes.

This suggests the future of practical and ethical AI lies in thoughtful, accountable human-machine partnerships, with new research focus shifting from optimizing algorithms, to designing interactions into, integrating into, and governing AI systems (Kirsten et al., 2025; Natarajan et al., 2024; Mayer et al., 2024). But with clearer paradigms, rigorous evaluation and steady, inclusive oversight, human–AI collaboration can be both effective and ethically defensible.

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). The 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 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
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. 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
8. 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
9. 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.
10. 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

11. iDigitalStrategies. (2024). Unraveling the quandary: The problem with AI-generated content detectors. https://www.idigitalstrategies.com/blog/problem-with-ai-generated-content-detectors/
12. Jang, J. (2024, June 10). Some thoughts on human-AI relationships. OpenAI Community. https://community.openai.com/t/some-thoughts-on-human-ai-relationships/1279464
13. 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.
14. 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
15. 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
16. 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
17. Kumar, V., Talwar, S., & Doshi, P. (2024). The diversity–innovation paradox in generative AI. arXiv. https://doi.org/10.48550/arXiv.2405.13868
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, 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
20. Mayer, L. W., & Karny, S. (2025). Human-AI collaboration: Trade-offs between performance and preferences. arXiv. https://doi.org/10.48550/arXiv.2503.00248
21. 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
22. 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
23. Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Science, 381(6654), 187–192.
24. 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
25. 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
26. 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/

27. 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
28. 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
29. Savage, T. (2023). Human-in-the-loop problem-solving with artificial intelligence. Academy of Management Review. https://doi.org/10.5465/amr.2021.0421
30. 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.
31. 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
32. SoK: Content moderation in social media, from guidelines to enforcement and research to practice. (2023). In 2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P) (pp. 488–506). IEEE. https://doi.org/10.1109/eurosp57164.2023.00056
33. 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
34. Wang, X., Chen, X., & Qu, Y. (2023). Constructing ethical AI based on the "human-in-the-loop" system. Systems, 11(11), Article 548. https://doi.org/10.3390/systems11110548
35. 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/
36. 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 Science.