

Ethical Boundaries in Human-AI Collaborative Decision Systems: A Review

Kumkum Suryavanshi, Department of Computer Science & Engineering, Indore Institute of Science & Technology, Indore, Madhya Pradesh, India kumkum.suryavanshi@indoreinstitute.com

Khushi Chouhan, Department of Computer Science & Engineering, Indore Institute of Science & Technology, Indore, Madhya Pradesh, India khushi.chouhan@indoreinstitute.com

Kumkum Shukla, Department of Computer Science & Engineering, Indore Institute of Science & Technology, Indore, Madhya Pradesh, India kumkum.shukla@indoreinstitute.com

Harsh Prajapat, Department of Computer Science & Engineering, Indore Institute of Science & Technology, Indore, Madhya Pradesh, India harsh.prajapat@indoreinstitute.com

Praveena Joshi, Department of Computer Science & Engineering, Indore Institute of Science & Technology, Indore, Madhya Pradesh, India praveena.joshi@indoreinstitute.com

ABSTRACT - This review covers human-artificial intelligence collaboration in making decisions, observing how scholarship has moved from thinking of AI as a passive tool to considering it as an active actor in healthcare, management, and ethics. While AI systems provide for efficient data processing and pattern recognition, humans provide contextual understanding, ethical judgment, creativity, and accountability. Previous literature suggests that well-designed collaboration between humans and AI may outperform completely automated decision-making. However, several key challenges persist, like bias, lack of transparency, mis calibrated trust, poorly defined distribution of roles, and accountability gaps. To handle these concerns, human-centered design, explainability in AI, well-articulated guidelines on ethical consideration, and mechanisms that effectively introduce oversight remain at the core of literature. These approaches jointly try to ensure that human-AI decision-making systems are fair, accountable, and trustworthy in high-stakes and critical settings.

INTRODUCTION

This work centers on the growing area of human-AI collaboration in decision-making processes. Human-AI collaboration can be seen to refer to collaboration between humans and AI systems when complementary strengths from both elements come together to develop superior outcomes. Indeed, recent studies show a clear shift away from considering AI as a passive tool towards understanding it as an active decision maker [1,5].

Drawing on literature from management, healthcare, autonomous systems, and ethics, this review interrogates how responsibilities are allocated between humans and AI within collaborative decision-making architectures [3,9]. In addition, while AI systems are good at the fast analysis of data, pattern recognition, and predictive tasks, human agents bring contextual understanding, ethical judgment, and responsibility for ultimate decisions [8,12]. However, ethical challenges such as algorithmic bias, lack of transparency, mis calibrated trust, and ill-defined boundaries of responsibility stand in the way of effectively deploying human-AI decision systems, especially in high-stakes domains [6,10].

This review synthesizes existing scholarship to analyze these ethical challenges through the dimensions of accountability, fairness, transparency, and human control. In integrating findings across disciplines, the paper will shed light on current limitations, identify gaps in the extant research, and propose directions for responsible development of trustworthy and ethically bounded human-AI decision-making systems.

LITERATURE REVIEW

Scholarship on human-AI collaboration increasingly reframes AI from a passive decision-support tool to an active participant in collaborative decision-making. Foundational conceptual analyses point out that effective collaboration arises when human judgment and AI capabilities are deliberately integrated rather than substituted for one another [1,5]. These studies emphasize complementary strengths: AI contributes speed, scalability, and pattern recognition while humans bring contextual reasoning, ethical awareness, and accountability for outcomes. More recent syntheses reinforce this perspective by outlining human-AI collaboration as a socio-technical system rather than a purely technical solution [9,13].

A few of these studies focus particularly on ethical decision-making within human-AI teams. Studies in autonomous systems and ethical modeling highlight that, when AI systems impact pivotal decisions, shared responsibility and moral oversight become paramount considerations [2,7]. Empirical and conceptual work indicates that ethical risk arises when decision authority becomes ambiguously divided between humans and AI, engendering accountability gaps [1,12]. From this perspective, AI ethics literature calls for clear human roles and constrained, ethical AI behavior in order to minimize unintended harm [11].

Another central theme in the literature is trust and reliance. Experimental comparisons between human and AI experts suggest that human–AI teams can outperform individual decision-makers, but only if trust is well calibrated [6,10]. Over-reliance on AI recommendations can produce automation bias; under-reliance can prevent teams from capitalizing on AI’s analytic strengths.

Research in medical and other high-consequence decision domains has shown that transparency and explainability have a profound impact on trust and effective collaboration [15]. More recent work broadens the perspective beyond dyadic human–AI interaction to take up governance and system-level issues. Reviews and framework-oriented articles identify fairness, transparency, accountability, and human control as mutually interdependent ethical dimensions that need to be addressed together rather than separately [9,13]. These analyses conclude that ethical human–AI collaboration requires integrated design strategies that intertwine technical safeguards with organizational practices and regulatory oversight. Overall, the literature suggests that, while human–AI collaboration carries great potential, its ethical feasibility depends on thoughtful system design, clear role specification, and continued human participation.

METHODOLOGY

This review uses a qualitative approach, combining conceptual analysis, a structured look at past research, and a summary of recent studies on making. The goal is to see how researchers define human–AI teamwork, what methods they use, and how these choices affect our understanding of trust, responsibility, and the quality of decisions made together.

Paper Selection and Scope: The studies reviewed include peer-reviewed journal articles, conference papers, and select preprints that focus on human–AI collaboration, ethical questions, and AI’s role in important decisions. Only research that looked specifically at how humans and AI interact, or share roles, was included. Papers on fully independent AI systems were left out. The selection covers both theoretical discussions and real-world tests of decision-support systems.

Conceptual and Theoretical Analysis: First, the review looked at conceptual and theory-based papers to see how they model humans and AI how humans and AI work together in decision-making, as partners in decision-making. These studies were checked for how they talk about human control, shared responsibility, ways to supervise AI, and the ethical rules for working together. Important ideas like agency, accountability, and oversight were pulled out to build a foundation for the rest of the review.

Analysis of Empirical Study Designs: Empirical studies were then examined for how they set up their experiments, what decision tasks they used, who participated, and how they measured results. Special attention was given to research comparing human experts with AI, or looking at different levels of AI involvement, like just giving advice versus making decisions. The review focused on how these studies measured trust, reliance, and responsibility and performance to see how well human-AI teams work together.

Cross-Study Synthesis: Finally, the review brought together findings from all the different types of studies to identify where they agree or disagree about ethical human–AI collaboration. Conceptual models of shared roles were compared with real-world evidence on trust and responsibility. This helped highlight both common trends and gaps in what we know, pointing to what future research should focus on for more ethical and effective human–AI partnerships.

ETHICAL CHALLENGES IN HUMAN–AI DECISION MAKING

When people and AI work together to make important choices, several ethical challenges tend to come up again. These include accountability (who’s responsible), bias, and Fairness, transparency (how decisions are made), and keeping meaningful human control. These issues are closely linked—if one is weak, it often makes the others worse. For AI-supported decisions to be trusted and accepted by society, each challenge needs to be addressed through both smart technical design and strict rules or oversight.

Accountability: To fix this, it’s important to clearly define everyone’s roles throughout the AI lifecycle, keep detailed records of decisions, and make sure there are ways to check what happened if problems arise. Laws and regulations also need to spell out who is liable and who oversees these systems as AI becomes more involved in decision-making.

Bias and Fairness: AI systems can sometimes reinforce unfairness if they are trained on biased data. This can lead to discrimination in areas like hiring, loans, healthcare, or criminal justice. When people and AI work together, there’s a risk that humans might just go along with AI suggestions—even if those suggestions are biased. Fighting

bias goes beyond just fixing problems after they show up. It means collecting data carefully, measuring fairness, and checking results regularly. It also requires including different voices in the design process to make sure certain groups aren't treated unfairly.

Transparency: Trust in AI decisions depends on understanding how those decisions are made. Many modern AI One major question is who should take responsibility when things go wrong. If an AI system plays a big role in a bad outcome, it can be unclear whether blame should fall on the designers, the people using the system, the company that owns it, or even government regulators. Because humans and AI share decisions, traditional ideas about responsibility don't always fit. Systems work in complex ways that are hard for people to understand—sometimes called “black boxes.” This makes it difficult to spot mistakes, challenge unfair decisions, or know if the system is working ethically. To improve transparency, AI systems should include ways to explain their decisions, provide clear documentation, and offer access to the data and logic behind their choices. Explanations should be easy to understand and relevant to the people affected by the decisions.

Human Control: It's vital to make sure that humans stay in control of decisions, especially when those decisions are sensitive or have big consequences. As AI gets more advanced, there's a risk that important choices could be handed over to machines without enough human input. Good human control means having clear points where humans can step in, making sure people and AI trust each other appropriately, and giving humans real power to change or stop AI actions if needed. People shouldn't just be figureheads—they need to stay engaged and informed throughout the process.

RESEARCH GAPS AND FUTURE DIRECTIONS

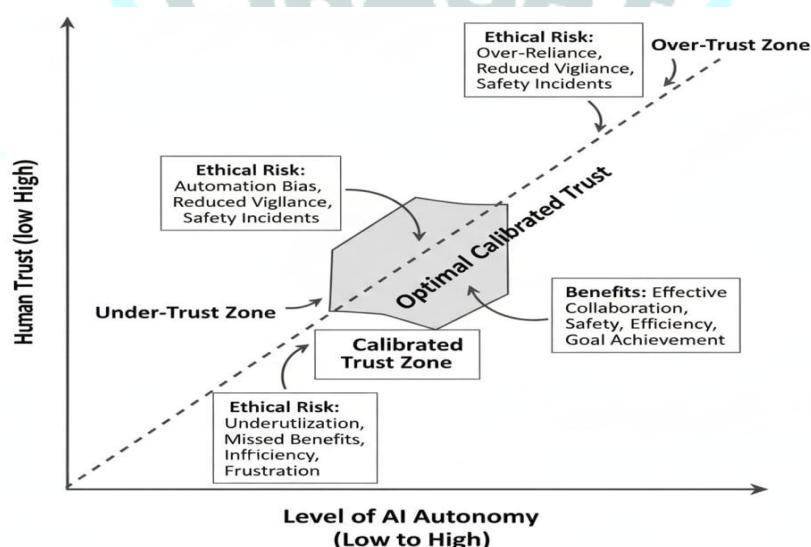
Although research on ethical human–AI decision systems is growing fast, there are still important gaps that need to be addressed to make these systems truly effective and trustworthy in the real world.

Limited Real-World Evidence in High-Stakes Situations

Much of what we know comes from lab experiments and simple decision tasks, not from real-life situations where the stakes are high—like healthcare, law, finance, or public policy. These studies often miss important factors such as workplace pressure, actual accountability, time limits, and legal risks. Moving forward, researchers should focus on studies that use real-world settings, long-term observations, and tests within organisations. This would help us understand how well accountability, fairness, transparency, and human control hold up when it really matters.

Simple Interaction Models: Even though we talk about “collaboration,” most current human–AI systems only provide one-off Studies show that just adding AI to the mix doesn't automatically make teams better. Problems often come from misplaced trust, over-reliance, or not understanding what AI can and can't do.

recommendations or basic advice. True teamwork—where people and AI discuss, challenge each other, and solve problems together—is still rare. Future research should try out richer ways for humans and AI to interact, such as step-by-step collaboration, resolving disagreements, or co-creating solutions. We also need better ways to tell the difference between basic assistance and real teamwork.



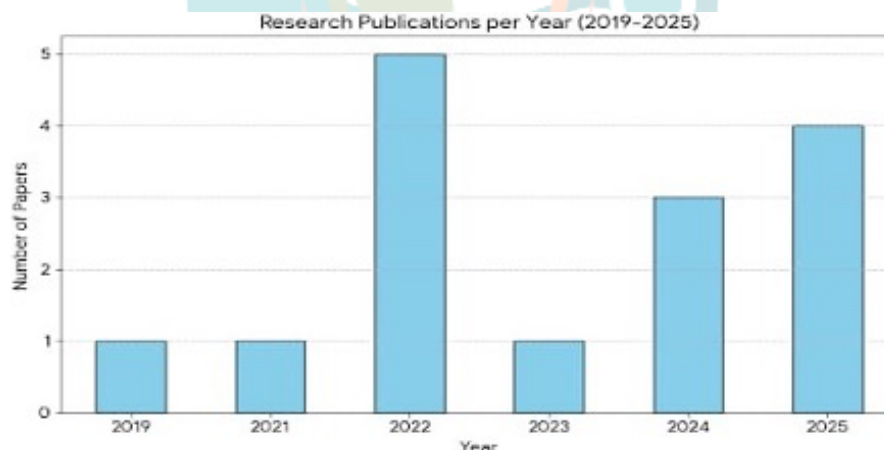
Fragmented Approach to Ethics: Ethical issues like fairness, accountability, and transparency are often studied separately, which means systems might be good in one area but weak in others. This piecemeal approach doesn't align with how these issues are interconnected in real life. A key step is to build methods that evaluate all these ethical aspects together. This way, we can see how fairness, clear responsibility, openness, and human control work side by side in real-world systems.

Human Factors and Trust: Future work should focus on helping humans build the right level of trust in AI, designing training and interfaces that show when to trust the system and when to question it.

Missing Socio-Legal and Global Perspectives: Most research is centered in a few countries and doesn't always consider global perspectives, especially those from the Global South or marginalized communities. This can lead to ethical guidelines that don't fit everywhere and sometimes reinforce existing inequalities. The field needs more research that compares different countries, includes voices from underrepresented groups, and develops flexible rules that can be adapted locally. This will help make human-AI collaboration fair and just for everyone.

Moral Reasoning and Value Alignment: Finally, there's still a big gap between technical work on human-AI teaming and deeper questions about values and moral responsibility. Many systems assume ethics can be captured by simple rules or numbers, without considering what decisions should—and shouldn't—be handed over to AI. Future research should bring together ideas from philosophy, ethics, and design to make sure humans stay meaningfully involved, especially for tough moral choices. Bridging this gap is key to setting boundaries and keeping AI aligned with what we value as a society.

LONGITUDINAL DISTRIBUTION OF RESEARCH LITERATURE (2019-2025)



YEARS OF RESEARCH PAPER	NUMBER OF PAPERS	CORE RESEARCH FOCUS AREA
2019	1	Technical foundations and complex reasoning
2021	1	Anticipatory Governance and Ethical Tools
2022	5	Human-AI Teaming and Ethical Risks Standards
2023	1	Generative AI Impact and Ethical Risks
2024	3	Organisational Productivity and Team Dynamics
2025	4	Agentic AI Systems and Governance Security

CONCLUSION

This review examined ethical boundaries in human-AI decision systems by focusing on accountability, bias and fairness, transparency, and human control. Across existing studies, it becomes clear that ethical issues are not secondary concerns but central factors that shape trust and responsibility in AI-supported decision-making, especially in high-stakes contexts. Future research must move toward realistic deployments, richer forms of human-AI interaction, and integrated ethical evaluation frameworks. Ultimately, the ethical success of human-AI decision systems depend not on AI sophistication alone, but on how effectively humans remain informed, accountable, and able to intervene.

REFERENCES

- [1] D.-T. Hoang and H.-J. Kang, "A survey on deep learning based bearing fault diagnosis," *Neurocomputing*, vol. 335, pp. 327–335, Mar. 2019, doi: 10.1016/j.neucom.2018.06.078.
- [2] Z. Zhu, Y. Lei, G. Qi, Y. Chai, N. Mazur, Y. An, and X. Huang, "A review of the application of deep learning in intelligent fault diagnosis of rotating machinery," *Measurement*, vol. 206, Art. no. 112346, Jan. 2023, doi: 10.1016/j.measurement.2022.112346.
- [3] M. Hakim, A. A. B. Omran, A. N. Ahmed, M. Al-Waily, and A. Abdellatif, "A systematic review of rolling bearing fault diagnoses based on deep learning and transfer learning: Taxonomy, overview, application, open challenges, weaknesses and recommendations," *Ain Shams Engineering Journal*, vol. 14, no. 4, Art. no. 101945, Apr. 2023, doi: 10.1016/j.asej.2022.101945.
- [4] X. Chen, R. Yang, Y. Xue, M. Huang, R. Ferrero, and Z. Wang, "Deep transfer learning for bearing fault diagnosis: A systematic review since 2016," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, 2023, doi: 10.1109/TIM.2023.3244237.
- [5] S. Tang, J. Ma, Z. Yan, Y. Zhu, and B. C. Khoo, "Deep transfer learning strategy in intelligent fault diagnosis of rotating machinery," *Engineering Applications of Artificial Intelligence*, vol. 134, Art. no. 108678, Aug. 2024, doi: 10.1016/j.engappai.2024.108678.
- [6] "Deep learning aided data-driven fault diagnosis of rotatory machine: A comprehensive review," *Energies*, vol. 14, no. 16, Art. no. 5150, 2021, doi: 10.3390/en14165150.
- [7] "From theory to industry: A survey of deep learning-enabled bearing fault diagnosis in complex environments," *Engineering Applications of Artificial Intelligence*, 2025, Art. no. 113068, doi: 10.1016/j.engappai.2025.113068.
- [8] "Deep learning-based fault diagnosis in wind turbine bearings and gearboxes using vibration signals: Survey, challenges, and recommendations," *IEEE Access*, vol. 13, pp. 207013–207032, Nov. 2025, doi: 10.1109/ACCESS.2025.3636831.
- [9] "A review of deep learning in rotating machinery fault diagnosis and its prospects for port applications," *Applied Sciences*, vol. 15, no. 21, Art. no. 11303, Oct. 2025.
- [10] R. Rony et al., "Deep learning and few-shot learning-enabled machine health monitoring for rotating machinery: A critical review of vibration-based fault diagnosis, prognostics, and industrial deployment," *Quality and Reliability Engineering International*, Jul. 2026, doi: 10.1002/qre.70302.
- [11] S. Pagare and R. Kumar, "Optimizing activity classification through bi-directional LSTM in human activity recognition systems," *Asian Food Science Journal*, vol. 6, no. 10, pp. 5838–5866, 2024. [Online]. Available: <https://doi.org/10.48047/AFJBS.6.10.2024.5838-5866>
- [12] S. Pagare and R. Kumar, "Human action recognition using long short-term memory and convolutional neural network model," *International Journal of Soft Computing and Engineering*, vol. 14, no. 2, pp. 20–26, Jun. 2024. [Online]. Available: <https://doi.org/10.35940/IJSCE.I9697.14020524>
- [13] S. Pagare, R. Kumar, and S. K. Gupta, "An adaptive hybrid deep learning approach for human action recognition," 2025, pp. 323–334. [Online]. Available: https://doi.org/10.2991/978-94-6463-716-8_26
- [14] S. Pagare and R. Kumar, "Object detection algorithms compression CNN, YOLO and SSD," 2023.
- [15] P. Shreyas, K. Rakesh, and S. K. Gupta, "Attention-driven BI-LSTM for robust human activity recognition and classification in disaster scenarios," *Disaster Advances*, vol. 18, no. 8, pp. 19–32, Aug. 2025. [Online]. Available: <https://doi.org/10.25303/188DA019032>

