Advances in Consumer Research
https://acr-journal.com/
Volume-3 | Issue-2 | Feb 2026
Original Researcher Article

Trust, Fairness and Safety in Ethical Artificial Intelligence: A Stakeholder Theory–Based Conceptual Framework

Dr.Shruthi J Mayur [1]
[1]Associate Professor, Department of HR & OB, T A Pai Management Institute, Manipal Academy of Higher Education, Manipal

ABSTRACT
The rapid diffusion of artificial intelligence (AI) and data-driven decision systems has fundamentally reshaped organizational processes, managerial judgment, and stakeholder relationships. While technical performance metrics and legal regulations exist, there remains a disconnect between rigorous engineering standards and the subjective expectations of diverse stakeholders. This paper proposes a novel conceptual framework that positions trust as the operational foundation of ethical AI, grounded in Stakeholder Theory. Drawing on stakeholder theory, we argue that ethical AI rests on two foundational dimensions, namely, fairness and safety, which reflect an organization's moral obligations toward its stakeholders. We conceptualize fairness as the absence of bias, discrimination, and systematic exclusion in data and algorithms, and safety as the protection of privacy, confidentiality, and security across the AI lifecycle. The intersection of these dimensions yields four distinct trust conditions that shape stakeholder acceptance, resistance, or disengagement. By integrating ethics, trust, and stakeholder theory, this paper advances a unifying conceptual model that clarifies how ethical AI generates legitimacy and sustained value. We conclude by outlining practical implications for organizational governance and proposing a future research agenda for management scholars....
Keywords: Ethical AI, Trust, Stakeholder Theory, Fairness, Data Governance, Algorithmic Ethics

INTRODUCTION:

Ethical use of data and AI hinges on trust. The integration of AI into corporate governance, healthcare, and national security has fundamentally altered the landscape of decision-making, shifting agency from human actors to algorithmic systems. Modern AI systems raise concerns about bias, privacy breaches, and opaque decision-making, all of which threaten stakeholder confidence. As these technologies become more autonomous, the reliance on legal frameworks to enforce ethical behavior has proven necessary but insufficient (Mirishli, 2025). Traditional governance structures strive to promote transparency and accountability, yet they often struggle to keep pace with the speed of technological advancement, particularly regarding highly advanced Large Language Models (LLMs) that may soon surpass human intelligence (Hossain & Ahmed, 2023). Consequently, the central problem facing the field is not merely technical competence, but the establishment of trust, a state where stakeholders can confidently rely on an AI system to act within ethical boundaries under uncertainty. From a stakeholder theory perspective, different groups have distinct vulnerabilities and expectations, and trust must be earned in relation to all of them.

While regulators explicitly demand that AI systems be fair and safe, end-users often seek explainability to develop trust. In this paper we argue that stakeholder trust in data and AI is the foundation of ethical AI use. We also propose that trust itself emerges from two orthogonal dimensions namely, fairness which includes unbiased, just algorithmic outcomes and safety which includes security, privacy protection, and responsible use. These dimensions combine to yield four distinct trust conditions. We develop this framework using stakeholder theory as our guiding lens, engage deeply with trust and ethics literature, and outline both theoretical contributions and practical implications. We also propose a detailed research agenda. The paper is organized as follows: the next section reviews stakeholder theory and trust in technology; we then present our fairness and safety trust model and describe its four conditions; we conclude with implications and future research directions.

Theoretical Background

Stakeholder Theory and AI Ethics

Stakeholder theory holds that organizations have moral and strategic obligations to all parties affected by their actions, not just shareholders . In the context of AI, this means firms must consider the interests of customers, employees, regulators, communities, and other stakeholders when designing and deploying AI systems. Recent work emphasizes that "AI systems change human interactions and ethical norms," requiring stakeholders to work together "to prevent AI from undermining human values and social cohesion". For instance, developers must identify passive stakeholders that is those affected by AI decisions but without direct power, and engage their representatives, so that AI development is morally and socially sustainable **(Brodny & Tutak, 2025; Miller, 2022)**. From a stakeholder perspective, trust in AI is

situation specific. An organization may be trusted by customers but distrusted by investors, depending on how it addresses each group's concerns. In fact, trust antecedents differ across stakeholder groups — employees may trust an AI system because management appears benevolent toward workers, whereas external stakeholders (like clients or regulators) may withhold trust if they perceive incompetence or risk. Thus, managing ethical AI involves balancing multiple stakeholder expectations for fairness and safety.

Trust in Technology and Organizations

Trust broadly means the willingness to be vulnerable to the actions of another based on positive expectations. In organizations, trust has three classic bases namely, ability (competence), benevolence (alignment with one's interests), and integrity (fairness and honesty). Importantly, trust in AI has both a systemic and institutional side. Stakeholders may trust a technology's performance while still distrusting the organization using it, and vice versa. Prior research indicates that initial trust in new systems is heavily influenced by the sponsoring agency's reputation and situational cues (Li et al., 2008). Hence institution-based and technology-based trust both are relevant. The stakeholder lens implies that different groups will weigh these factors differently: regulators might focus on systemic safety and fairness, whereas end-users might focus on clarity and experience. This distinction between trust in the AI system versus trust in the deploying organization echoes findings in trust and IS research, where perceptions of institutional support and social norms influence trust just as much as the artifact's features.

Finally, trust is dynamic. In human AI interactions, initial trust in AI is often low, especially for novel or opaque systems. Users tend to begin with caution until the system proves itself. Studies of AI infused robots and decision aids find that transparency and reliability are key to building trust over time (Glikson & Woolley, 2021). In practice, trust may evolve differently across domains. People often start with higher trust in easily observable entities and lower trust in opaque algorithms, but this can reverse as outcomes accumulate. But once violated by a biased outcome or data breach, trust in AI is difficult to restore. Human AI trust repair draws on strategies from social psychology that state apologies and explanations can modestly help, but technical fixes (e.g. updating the model) tend to be more effective in rebuilding trust (Pareek et al., 2024). These dynamics emphasize that ethical AI systems must not only start "trusted" but maintain and, if necessary, repair that trust through transparent and responsible practices.

Fairness and Safety as Ethical Dimensions

Two ethical dimensions underpin trust in AI namely fairness and safety. Fairness refers to unbiased, just treatment of stakeholders in data and algorithmic processes **Chen et al. (2025)**. In traditional stakeholder and justice theory, perceived fairness in procedures and distributions is foundational for trust **(Colquitt & Rodell, 2011)**. For AI, fairness issues include algorithmic bias against protected groups and opaque decision rules. Experimental studies show that when people perceive algorithmic offers as fair, they invest more trust, whereas unfair AI behaviours elicit distrust. In fact, Chen et al. (2025) find participants who received fair offers had higher trust investment from an AI compared to unfair offers. Conversely, unfair or discriminatory AI outcomes directly erode confidence. Meta-analytic research in organizations likewise finds that informational fairness (clear, truthful explanations) strongly predicts subsequent trust in authorities. In short, perceived fairness (or justice) in AI decisions is a critical precursor to trust.

Safety encompasses system security, data privacy, robustness, and responsible usage practices. A safe AI system is one that guards stakeholder data, resists adversarial attacks, and limits harm. Safety concerns cut to the core of stakeholder vulnerability and data breaches or malevolent use of AI violate social contracts and destroy trust. Official guidelines pair technical robustness and safety with fairness as pillars of trust. AI governance frameworks highlight that robustness, reliability, and resilience against manipulation are fundamental to stakeholder confidence (Mitra, 2025). A safe AI system is one that stakeholders can rely on without fear of security lapses or privacy violations. Organizational literature suggests stakeholders readily trust organizations perceived as responsible with data and security. Conversely, any compromise can quickly weaken trust. In practice, leaders emphasize privacy preserving design and secure by design principles so that innovation aligns with societal values and regulatory expectations.

Taken together, fairness and safety form two independent axes. Fairness ensures just treatment of stakeholders, while safety ensures protection from harm. We argue that stakeholder trust in AI arises when both dimensions are sufficiently addressed. If either dimension is lacking, trust is impaired even if the other is present. Below we formalize these ideas in a conceptual model with four trust conditions.

A Conceptual Model of Trust, Fairness, and Safety

We propose a 2×2 trust model with fairness and safety as orthogonal dimensions. Each dimension is binary (high/low) for conceptual clarity, producing four trust conditions as given in Figure 1.

High Fairness, High Safety (Full Trust): When AI systems are both fair and safe, stakeholders can fully trust the outcomes. Fairness assures stakeholders that the system treats them justly, while safety ensures that the system is secure, private, and robust. In this ideal quadrant, organizational and system trust align stakeholders have no reason to suspect bias or danger. Our theory predicts *strong trust*: stakeholders are willing to be highly vulnerable to the AI or organization, leading to enthusiastic adoption and compliance.

High Fairness, Low Safety (Fair but cautious trust): Here the AI appears fair in its decisions without evident bias or unequal treatment. But it has unresolved safety issues such as security vulnerabilities or privacy risks. Stakeholders may trust the system's benevolence and integrity knowing outcomes are fair, but they worry that data breaches or malfunctions could occur. We call this Conditional Trust. Stakeholders grant trust conditionally

on safe operation.

Low Fairness, High Safety (Safe but skeptical trust): In this quadrant, the AI is technically robust and protected, but stakeholders perceive unfairness in its outputs. Stakeholders may acknowledge the system's competence by trusting its security and accuracy but question its integrity and goodwill. We term this Skeptical Trust. Stakeholders rely on the system's performance but remain skeptical about ethical intent.

Low Fairness, Low Safety (No trust): When an AI is both unfair and unsafe, trust collapses. Stakeholders see neither justice nor protection. They feel outcomes are biased and the system is vulnerable. This quadrant yields explicit distrust or rejection. Stakeholders in this condition withdraw trust immediately and may demand external oversight or halt adoption.

These four conditions map to stakeholder experiences. Empirical work supports this conception. For instance, participants in trust games invest more in fair proposers than unfair ones, illustrating that fairness alone boosts trust. Conversely, when unfairness is perceived, users often trust an AI more than a human precisely because they view the AI as lacking malicious intent, a nuance we incorporate by noting that even in an unsafe AI, stakeholders might place *some* trust if they believe bias is not an issue. Similarly, stakeholders tend to trust stable, reliable systems (safety) unless fairness norms are violated. Our model thus synthesizes these insights: trust is highest only when both justice and safety are satisfied.

Figure 1. A conceptual model of trust in AI

HIGH		
FAIRNESS	**FULL TRUST** High Fairness and High Safety. Ethical and Secure AI	**FAIR BUT CAUTIOUS TRUST** High Fairness but Low Safety. Fair but insecure AI
	SAFE BUT SKEPTICAL TRUST Low Fairness and High Safety. Secure but unfair AI.	**NO TRUST** Low Fairness and Low Safety. Unfair and Risky AI
LOW	**SAFETY**	**HIGH**

Beyond individuals, Stakeholder Theory refines this model. Different stakeholders prioritize these dimensions differently. Regulators (concerned with systemic risk) may view any safety lapse as unacceptable, even if fairness holds, pushing trust toward the cautious doubtful zones. On the contrary, community groups focused on equity will view any fairness lapse as trust breaking, even if the safety element is high. Thus, any practical ethics framework must navigate these stakeholder-specific thresholds.

Implications

Theoretical Contributions.

This framework contributes to stakeholder and trust literatures in several ways. First, it explicitly integrates Stakeholder Theory with AI ethics, showing how stakeholder expectations of fairness and safety co-create trust. While previous work acknowledges multi stakeholder views of ethical AI, our model formalizes trust as the central emergent outcome. Second, we clarify the multi-dimensional nature of trust in AI. Rather than a single continuum, trust arises from the intersection of moral (fairness) and technical (safety) considerations. This bridges organizational justice research, where fairness predicts trust, with security and privacy research, where system robustness emphasizes trust. Third, by naming four distinct trust conditions, this paper offers a nuanced taxonomy for future theory. Existing studies often dichotomize trust as trusted vs distrusted. However, our model suggests gradations such as conditional vs skeptical trust, which can inform more precise theorizing. Finally, our stakeholder lens highlights that trust is not monolithic. The same AI system can inhabit different trust conditions for different stakeholder groups. This advances stakeholder theory by mapping specific ethical dimensions to stakeholder vulnerabilities, showing that trust is situation specific.

Managerial and Practical Implications.

For practitioners and policymakers, our framework highlights that ethical AI is trustful AI. Organizations must address both fairness and safety proactively. In practice, this means implementing bias mitigation, transparency, and fairness in audits and rigorous security, privacy safeguards, and oversight. For instance, compliance teams should enforce data hygiene and robust encryption (safety) while development teams should ensure algorithms that are tested on diverse datasets (fairness). When either dimension is weak, managers should anticipate stakeholder wariness. For example, if an AI passes fairness test but has security vulnerabilities, product managers must communicate safeguards or restrict use to avoid eroding trust. Conversely, if safety is high but fairness is questioned, businesses should engage stakeholders transparently or halt biased features.

From a governance standpoint, stakeholders must be involved in trust building. Consistent with Stakeholder Theory, firms should consult regulators, customers, and civil society when defining fairness and safety standards. Trust-building strategies include clear communication of how data are protected and why decisions are fair. For instance, publishing algorithmic impact assessments and obtaining third-party audits can signal both competence (safety) and integrity (fairness). If trust is breached, explicit trust repair actions are needed. In such cases apologies and explanations help address the fairness dimension, while rapid technical fixes address safety. For high-stakes AI use in healthcare, finance, etc., embedding human oversight is critical to maintain trust. It should be

noted that autonomy plus accountability improves both fairness and safety perceptions.

Future Research

This framework opens numerous research directions. Empirical testing of the model is a priority (Shaiken et al., 2024). Researchers can design surveys and experiments to measure stakeholders perceived fairness and safety of an AI system and observe corresponding trust behaviours (Colquitt et al., 2001). Longitudinal studies could track how trust moves between our four conditions as AI systems evolve or as information emerges. Comparative studies across stakeholder groups could clarify whether regulators and customers' trust levels converge or diverge under various fairness/safety scenarios? The conceptual work should refine the nature of each trust condition.

Another area for further research is measurement. Developing validated scales for AI fairness perception and AI safety perception will help quantify dimensions identified in our model. This links to existing organizational justice measures (Colquitt et al., 2001) and trust in technology scales. Further, researchers can also explore contextual factors such as, how culture, industry norms, or regulatory environments shift the importance of fairness vs safety? Or how in highly regulated sectors, safety might dominate as a trust condition, whereas in consumer tech, perceived fairness could be paramount.

Finally, the model suggests interventions to study. How effective are different trust building actions in each condition? For example, when trust is low that is unfair or unsafe, are apologies or compensation enough? Or should the organization look at structural changes such as transparency or governance? The trust repair literature provides some insight, but more work is needed on algorithmic contexts. Crucially, given AI's rapid evolution, examining trust dynamics over extended use is vital. Some questions that need answers are: Do stakeholders learn to trust (or distrust) an AI more after repeated interactions? What roles do system explanations and user training play in calibrating trust? Each of these questions will deepen our understanding of how fairness and safety shape trustworthy AI.

CONCLUSION

In sum, we argue that trust is the cornerstone of ethical data and AI use, and that stakeholder trust arises from two interrelated pillars of fairness and safety. By applying stakeholder theory, we have shown that different groups demand both just and secure AI, and that meeting these demands generates trust. Our four-quadrant model clarifies the conditions under which trust flourishes or fails. This contribution has both theoretical and practical value as it extends stakeholder theory into AI ethics and offers managers a strategic lens for building trust. Finally, fostering ethical AI adoption means attending to the fairness of outcomes and the safety of processes, thus earning the confidence of all stakeholders. Future research should build on this framework to test and expand its implications, ensuring that trust and the ethical adoption of AI is sustainable in an increasingly data driven world...

.

REFERENCES

1. Brodny, J., & Tutak, M. (2025). Stakeholder interactions and ethical imperatives in big data and AI development. Journal of Open Innovation: Technology, Market, and Complexity, 11(1), 100491.
2. Chen, R., Jin, Y., Yu, L., Tempel, T., Li, P., Zhang, S., Li, A., & He, W. (2025). The influence of perceived fairness on trust in human-computer interaction. International Journal of Psychology, 60(5), e70111.
3. Colquitt, J. A., & Rodell, J. B. (2011). Justice, trust, and trustworthiness: A longitudinal analysis. Academy of Management Journal, 54(6), 1183–1206.
4. Glikson, E., & Woolley, A. W. (2021). Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 15(2), 655–688.
5. Li, X., Hess, T. J., & Valacich, J. S. (2008). Why do we trust new technology? A study of initial trust formation with organizational information systems. Journal of Strategic Information Systems, 17(1), 39–71.
6. Miller, G. J. (2022). Stakeholder roles in artificial intelligence projects. Project Leadership and Society, 3, 100068.
7. Mitra, M. (2025). Ethical theories, governance models, and strategic frameworks for responsible AI adoption and organizational success. Frontiers in Artificial Intelligence, 8, 1619029.
8. Pareek, S., Velloso, E., & Goncalves, J. (2024). Trust development and repair in AI-assisted decision-making during complementary expertise. In Proceedings of the ACM on Fairness, Accountability, and Transparency (FAccT 2024), 1–16.
9. Sarker, I. H., et al. (2026). SME-TEAM: Leveraging trust and ethics for secure and responsible use of AI and LLMs in SMEs. npj Artificial Intelligence, 2, 12.
10. Shaiken, I., et al. (2024). Trust in AI: Progress, challenges, and future directions. Humanities and Social Sciences Communications, 11, 987.

.