Human Factors in Artificial Intelligence and Social Computing, Vol. 203, 2026, 58–67
https://doi.org/10.54941/ahfe1007318

Trust and Calibration in AI-Mediated Decision Support Under Conditions of Risk

Angela M Fike, Tian Wang, and Masooda Bashir

University of Illinois, Urbana-Champaign, Urbana-Champaign, IL 61820, USA

ABSTRACT

AI-mediated decision support systems are increasingly deployed in domains characterized by risk, uncertainty, and time pressure. In such environments, appropriate reliance on AI recommendations requires not only initial trust formation but also dynamic recalibration when system performance fluctuates or conflicts with other information sources. Although determinants of perceived trust (e.g., explainability, authority cues, and ethical framing) have been widely studied, less attention has been given to how reliance behavior adjusts following observed system error. This paper presents a focused qualitative synthesis of empirical studies examining trust and reliance in AI-based decision support under conditions of risk or informational divergence. Across the included studies, trust was frequently operationalized as an attitudinal construct or predictor of adoption. In contrast, fewer investigations directly measured behavioral reliance following performance degradation or assessed calibration accuracy, defined as the alignment between perceived system capability and actual performance over time. Findings suggest that reductions in reported trust do not consistently translate into commensurate changes in reliance behavior. This divergence highlights the need to distinguish attitudinal trust from behavioral calibration when evaluating AI systems in safety-relevant contexts. We argue that calibration-aligned design (rather than trust maximization alone) should guide the development and assessment of high-stakes AI decision support.

Keywords: AI agent-interaction, AI-generated recommendations, Decision support, Trust calibration, High-risk scenario

INTRODUCTION

Artificial Intelligence (AI) agents are increasingly integrated into decision support systems across domains, including transportation, finance, and healthcare. In high-risk contexts in particular, AI-generated recommendations may influence time-sensitive decisions involving diagnosis, triage, treatment selection, or patient self-management. In such environments, errors carry significant consequences, and human decision makers must interpret AI outputs alongside other information sources, such as clinician guidance, prior knowledge, or personal experience. As AI agents become embedded within multi-source, time-pressured decision environments, understanding how humans calibrate trust and reliance in these systems becomes a central human factors challenge.

Received February 7, 2026; Revised April 7, 2026; Accepted April 21, 2026; Available online July 20, 2026

Trust has long been recognized as a key determinant of automation use, influencing whether users appropriately rely on, underutilize, or over-rely on automated systems (Lee & See, 2004; Hoff & Bashir, 2015; Parasuraman & Riley, 1997). Within this framework, inappropriate reliance has been described as a form of automation bias, in which users defer to system outputs even when contradictory evidence is available (Parasuraman & Riley, 1997). Appropriate reliance, therefore, depends on calibration, the alignment between perceived system capability and actual system performance. Miscalibration, in contrast, occurs when perceived capability exceeds or falls short of actual system performance.

However, much of the foundational trust-in-automation literature was developed in contexts involving structured tasks, relatively stable feedback loops, or single-source decision environments. AI agents in high-risk scenarios differ in important ways. For example, they often operate under uncertainty, produce probabilistic or generative outputs, and are deployed in settings where users must reconcile competing or conflicting information sources. Meanwhile, informational conflict is defined here as situations in which AI-generated advice diverges from other available evidence or expectations, introducing additional cognitive demands that may alter trust formation and adjustment processes. In high-stakes contexts, such conflict can interact with perceived authority, ethical framing, risk sensitivity, and time pressure, potentially amplifying miscalibration.

Despite these realities, within the sampled empirical studies, research has only begun to systematically examine how trust and reliance evolve under explicit AI-human disagreement and high-risk framing. Among the sampled empirical studies, research on trust in AI is limited in its manipulation of explicit AI-human disagreement and in its measurement of behavioral recalibration after conflict. Many studies assess perceived trust in isolation or focus on accuracy effects without incorporating behavioral reliance measures, high-risk framing, or longitudinal recalibration following error exposure. These gaps motivate a closer examination of how trust is conceptualized and operationalized in existing AI trust research, especially in high-risk scenarios (Parasuraman & Riley, 1997).

Therefore, in this literature review, we examine how empirical research on AI-mediated decision support conceptualizes and measures trust formation, behavioral reliance, and recalibration under conditions of risk or informational conflict. The objective was not to conduct a comprehensive systematic review of all AI trust research, but rather to evaluate how selected empirical studies operationalize dynamic trust adjustment in high-risk decision contexts.

RELATED WORKS

Trust in AI-mediated decision support has been conceptualized across multiple theoretical traditions, each offering distinct insight into how trust is formed, maintained, and repaired. Integrating these frameworks clarifies the distinction between initial trust formation and subsequent trust recalibration under conditions of error or informational conflict. Mayer, Davis, and Schoorman's

(1995) integrative model of organizational trust conceptualizes trust as a function of the trustee's perceived ability, benevolence, and integrity. Within AI-mediated contexts, these dimensions map onto perceptions of system competence (accuracy and performance reliability), alignment with user goals or ethical standards, and perceived legitimacy or institutional backing. Importantly, Mayer et al. distinguish between trust beliefs and risk-taking behavior in relationships (RTR), emphasizing that trust is not equivalent to behavioral reliance but rather a psychological state that precedes action under uncertainty. This distinction is critical for AI systems, in which users may continue to rely on decision support despite fluctuations in perceived trust, or, conversely, withhold reliance despite favourable attitudes.

Building on this foundation, Hoff and Bashir (2015) propose a multi-layer model of trust in automation consisting of dispositional trust (stable individual tendencies), situational trust (contextual influences such as task risk or time pressure), and learned trust (experience-based updating derived from system performance). This layered structure implies that trust calibration is not driven solely by immediate performance feedback but rather by the interaction among stable predispositions, contextual framing, and accumulated system experience. In high-risk environments, situational pressures and authority cues may moderate the extent to which performance degradation produces behavioral recalibration.

Trust repair research further complicates the performance-feedback assumption. De Visser, Pak, and Shaw (2018; 2020) demonstrate that trust erosion following automation failure does not necessarily mirror the rate or structure of trust acquisition. Trust repair may require different informational interventions (e.g., apologies, explanations, transparency signals) than those that initially generated trust. Moreover, users may attribute failures to transient external factors rather than systemic unreliability, slowing behavioral disengagement. This asymmetry between trust building and trust repair suggests that recalibration is not simply a linear function of observed accuracy.

These theoretical perspectives suggest that trust formation, trust adjustment, and behavioral calibration are related but distinct processes. Initial trust may be shaped by perceived competence, legitimacy, and the alignment of explanations. In contrast, recalibration depends on how users interpret errors, attribute causality, and integrate new evidence under contextual constraints. Additionally, trust repair can be viewed as a bilateral, dynamic process (Kim et al., 2009). This theoretical integration motivates the presented review's focus on whether empirical AI trust studies distinguish between attitudinal trust and behavioral reliance, and whether they systematically examine recalibration under conditions of risk and informational conflict, by coding studies by the construct measured (trust attitude, behavioral reliance, calibration alignment).

METHODS

Search Strategy A structured search was conducted in November 2025 using the Web of Science Core Collection. Search terms were constructed to capture research on AI-generated recommendations, trust, decision support, and

risk-related contexts. Boolean combinations included variations of: ("AI" OR "AI agent" OR "AI-generated recommendations") AND ("trust" OR "trust calibration" OR "reliance") AND ("decision support" OR "human decision making") AND ("risk" OR "high stakes" OR "conflicting information"). To narrow the scope to individual-level decision processes, studies focusing exclusively on clinical workflow optimization, retail systems, educational applications, or multi-device ecosystems were excluded. Review articles, commentaries, and purely theoretical papers were also excluded to prioritize empirical evidence.

Inclusion Criteria Studies were included if they: 1) Examined human interaction with AI-based decision support systems, 2) Measured trust, reliance, or trust calibration outcomes, 3) Involved experimentally manipulated or contextually framed risk, uncertainty, or informational divergence, and 4) Reported original empirical findings. Note that for our review, high-stakes contexts are defined as decision environments in which incorrect decisions may cause material, financial, or health-related harm (Parasuraman & Riley, 1997). The review process included two phases: 1) title and abstract screening, and 2) full-text review. Based on the selection criteria, a total of six peer-reviewed journal articles were included for further analysis. Note that our review is intended as a focused qualitative review rather than a comprehensive systematic review, emphasizing conceptual patterns in trust calibration measurement in high-risk scenarios.

Analytical Approach For each study, we extracted the decision context, the operationalization of trust (attitudinal vs. behavioral), manipulated variables (e.g., accuracy, explanation type, authority cues), the presence of error exposure, and evidence of reliance adjustment over time. We conducted a structured qualitative comparison to identify patterns in how trust formation, behavioral reliance, and recalibration were measured and interpreted.

FINDINGS

Trust as a Primary Driver of Reliance

In several included studies, perceived trust predicted adoption or continued use, even when performance cues were available (Choudhury & Shamszare, 2023; Klingbeil et al., 2024). Trust strongly predicted continued system use; most studies measured adoption or sustained engagement rather than behavioral recalibration following error. In at least one experimental decision-support setting, behavioral reliance declined more modestly than trust ratings following a reduction in accuracy (Klingbeil et al., 2024).

The reviewed studies also suggest that perceived trust may influence reliance behavior independently of objective system accuracy. This pattern reinforces foundational automation theory, suggesting that reliance depends on perceived capability rather than verified performance (Lee & See, 2004; Parasuraman & Riley, 1997). In AI-mediated environments, trust appears to function as a heuristic cue that simplifies decision-making under uncertainty. When users perceive AI systems as credible or authoritative, reliance may persist even when outputs are imperfect or probabilistic.

Table 1 synthesizes the contextual domains, trust determinants, and observed outcomes across the reviewed studies. First, trust determinants extend beyond system accuracy to include perceived usefulness, authority cues, ethical framing, transparency, and explanation quality. Second, trust outcomes are frequently operationalized as stated trust, adoption, or acceptance rather than observable recalibration following error exposure. Third, contextual moderators, such as risk framing and ethical positioning, shape perceived trust, but few studies directly examine their impact on behavioral reliance decisions. It's important to note that not all included studies involved real-world safety consequences; in several cases, risk was experimentally framed rather than materially consequential. This variation also reflects ongoing inconsistency in how 'high-risk" is operationalized across the included AI trust studies.

Table 1: Literature summary: context, trust determinants, and study observations.

Authors	Context	Primary Trust Determinants Examined	Trust Outcome or Observation
Choudhury & Shamszare (2023).	General Chat GPT	Trust, perceived usefulness & risk, transparency	Trust strongly predicts adoption and continued use; ethical concerns reduce trust
Shin (2021)	Simulated environmental decision context	Explainability, causality, perceived transparency,	Causality and explanation quality increase trust when aligned with mental models
Fahnenstich et al. (2024).	Risk-based decision support	Source of advice (human vs. AI), perceived risk	AI trust declines relative to human trust under high risks
Klingbeil et al. (2024).	Experimental AI decision support	Initial AI accuracy, authority cues, and feedback	Reliance declined more modestly than trust ratings following error exposure
Omrani et al. (2022).	AI systems across domains	Ethics, transparency, accountability, context	Trust is multidimensional and context sensitive
Wang et al. (2023).	Healthcare use of ChatGPT	Ethics, accountability, safety, and over-trust	Over-trust poses ethical and safety risks in healthcare

Most studies conceptualize trust as a multidimensional construct shaped by cognitive (explanation alignment), social (authority signalling), and contextual (risk, ethics) factors, rather than by accuracy alone. Also, trust outcomes are frequently operationalized as perceived trust, adoption, or acceptance rather than behavioral recalibration following error exposure. In addition, explicit manipulation of informational conflict is rare, and behavioral error experiments are limited within the sampled empirical studies. In the sampled empirical studies, trust-formation constructs were operationalized more frequently than dynamic trust-adjustment processes.

Explainability and Authority

The reviewed studies suggest that explainability is not uniformly effective in calibrating trust. Explanations appear to influence reliance primarily when

they are cognitively interpretable, task-relevant, and aligned with users' mental models. Prior experimental research supports this distinction. Kuang et al. (2020) demonstrate that explanation style affects user trust judgments, particularly when explanations align with users' expectations regarding system reasoning. Vasconcelos et al. (2023) further show that explanation fidelity can alter reliance behavior independently of objective model performance. These studies suggest that explanation effectiveness depends on psychological alignment rather than technical transparency alone. Overly complex, misaligned, or low-fidelity explanations may fail to promote appropriate reliance even when they increase perceived transparency.

Authority signaling effects may contribute to trust judgments in AI systems, though causal mechanisms remain underexplored. Perceived objectivity, computational expertise, or authority cues (as in the Klingbeil, Table 1) may amplify deference to AI recommendations. In high-risk healthcare contexts, such signals may unintentionally promote overreliance if not carefully calibrated. What stands out for designers is that they must consider not only what information is presented, but how it is framed and interpreted within broader sociotechnical contexts.

Trust Persistence and Recalibration Following Error

Table 2 summarizes how the reviewed studies addressed AI error exposure and subsequent trust adjustment. Across the included studies, trust was frequently operationalized as a self-reported attitude or as an intention to adopt. Fewer investigations measured behavioral reliance following performance degradation, and even fewer assessed calibration accuracy, whether reliance behavior proportionally tracked observed system performance over time, or explicitly examined trust repair dynamics following AI error.

Where error exposure was experimentally introduced, reported trust declined modestly; however, behavioral reliance did not consistently shift to the same extent as reported trust ratings. In other cases, reliance behavior was not directly measured following performance degradation, limiting conclusions about recalibration processes. Two studies documented perceived over trust or sustained engagement despite acknowledged inaccuracies, yet these observations were not consistently tied to longitudinal behavioral measures. To our knowledge, none of the reviewed studies employed formal calibration analysis (e.g., forecast reliability or probabilistic alignment metrics; Murphy & Winkler, 1977).

The reviewed evidence indicates that while trust formation is frequently operationalized and measured, systematic behavioral assessment of trust adjustment following error remains comparatively limited. As a result, conclusions regarding recalibration mechanisms must be interpreted cautiously. These observations reinforce the need for longitudinal and behavior-based research designs. Single-trial error manipulations provide limited insight into how reliance evolves over repeated interactions. To understand recalibration in practice, future studies must examine how individuals revise—or fail to revise—their reliance patterns across repeated interactions, especially in environments characterized by risk, ambiguity, and competing sources of authority.

Table 2: Trust breakdown and recalibration across reviewed studies.

Authors	Error Exposure	Observed Trust Change	Behavioral Recalibration and Notes on Calibration Pattern
Choudhury & Shamszare (2023).	No controlled error exposure (perception-based)	Trust influenced adoption and continued use	Not directly measured; Trust shaped sustained engagement rather than recalibration.
Shin (2021)	No explicit system errors, explanation manipulation not an error	Trust increased when explanations aligned with mental models	Not directly tested behaviorally; calibration is dependent on cognitive alignment, not performance
Fahnenstich et al. (2024).	Risk manipulation (not error)	Trust declined under high-risk framing relative to human advice	Partial behavioral shift; risk influenced stated trust more than actual reliance
Klingbeil et al. (2024).	Explicit AI errors are introduced; it explicitly manipulates the error	Trust declined modestly	Overreliance persisted despite errors, slow and incomplete trust erosion, with limited recalibration.
Omrani et al. (2022).	Perceived ethical/system concerns	Trust varied by ethical framing	Not behaviorally measured; trust calibration is conceptualized as multidimensional and context-sensitive.
Wang et al. (2023)	Inaccurate ChatGPT outputs; discusses inaccurate outputs, but does not test behavioral recalibration experimentally.	Acknowledged over-trust risks	Discusses potential overreliance risks but does not empirically test behavioral updating

Implications for Human Factors and AI System Design

These findings suggest that trust calibration (rather than trust maximization) should guide the design of AI systems in high-risk domains. Systems optimized solely to increase perceived trust or user acceptance risk amplification, thereby exacerbating overreliance when behavioral updating mechanisms are weak. Design evaluation should prioritize trust calibrated design, therefore move beyond attitudinal trust measures and assess whether reliance decisions appropriately track system capability across varying performance conditions and conflicting information. Design interventions such as calibrated confidence displays and explicit uncertainty communication align with established principles for supporting appropriate reliance in automation (Lee & See, 2004); Parasuraman & Riley, 1997). Similarly, structured disagreement prompts, and post-error reflection mechanisms are consistent with models emphasizing dynamic trust updating (Hoff & Bashir, 2015).

Designs that support calibration must clearly communicate uncertainty and system limits (Lee & See, 2004; Parasuraman & Riley, 1997). Approaches

such as calibrated confidence displays, explicit uncertainty communication, structured disagreement prompts, and post-error reflection mechanisms warrant systematic empirical comparison. This is needed not only for their effect on perceived credibility, but for their influence on observable reliance behavior under realistic stressors.

From a human factors perspective, this shifts design attention from one-time trust gains to how interfaces support ongoing reliance decisions as performance changes. Interfaces should help users recognize when system performance changes and prompt reconsideration of prior reliance patterns. In high-stakes environments, mis-calibrated trust is not merely a usability concern but a risk-management issue. Designing for appropriate reliance, therefore, requires interdisciplinary coordination across human factors engineering, domain expertise, and AI system development.

DISCUSSION

This review examined how empirical research on AI-mediated decision support conceptualizes and measures trust, reliance, and calibration in risky or uncertain contexts. Across the analysed studies, determinants of initial trust (e.g., explainability, authority cues, usefulness, and ethical framing) were represented more consistently than the mechanisms governing trust adjustment following error or informational conflict.

A consistent pattern across the reviewed studies is the gap between what users report and how they behave. Trust ratings often predicted adoption or continued use, yet few studies examined whether reliance shifted in step with changing system performance. This suggests that once credibility is established, reliance may become relatively stable, even as system accuracy fluctuates. Designing for appropriate use therefore requires more than influencing attitudes; it requires attention to how reliance decisions are made and revised in practice.

The reviewed literature also illustrates that trust dynamics are context-sensitive. In high-risk or time-pressured environments, perceived authority and ethical legitimacy may anchor user confidence more strongly than empirical performance cues. Users may reinterpret or discount occasional system failures, attributing them to situational complexity rather than system weakness. These attribution patterns complicate performance-feedback models of trust and imply that calibration is unlikely to be achieved reliably through exposure to errors alone. Instead, it requires continuous informational cues that clarify system boundaries, uncertainty, and intended use.

Research that effectively integrates behavioral data into trust studies will yield deeper insights into the challenges of trust calibration. Reliance should also be evaluated through measurable actions, such as acceptance rates, decision overrides, or time-to-decision, rather than self-reported attitudes alone. Longitudinal designs could capture how trust evolves across repeated interactions and changing performance conditions, providing a more realistic picture of recalibration in practice.

For designers and human-factors researchers, the findings underscore that the goal is not to maximize trust but to align it with actual system capability. Interfaces should make uncertainty explicit and invite critical engagement, particularly in high-stakes applications. Techniques such as confidence

visualization, uncertainty labelling, and transparent disagreement prompts may help users interpret recommendations more appropriately. Evaluating these mechanisms will require collaboration among engineers, behavioral scientists, and domain specialists to test how design interventions influence both perceptions and behavioral outcomes over time.

Overall, the literature provides insight into how trust in AI systems is formed, but far less clarity about how reliance changes once performance shifts or errors occur. Much of the evidence remains grounded in attitudinal measures, leaving behavioral adjustment comparatively underexamined. To better understand calibration in practice, future research should prioritize observable indicators of reliance across changing accuracy and risk conditions. Without such measures, conclusions about recalibration remain provisional. More precise behavioral evidence would allow stronger claims about how users respond to uncertainty, disagreement, and system failure over time.

CONCLUSION

This review highlights the need to distinguish between attitudinal trust and observable reliance when evaluating AI decision support under conditions of risk. Although determinants of initial trust are well documented, fewer studies have directly assessed how reliance behavior shifts as system performance changes. Future research should prioritize behavior-based measures and longitudinal designs that allow calibration processes to be observed across repeated interactions and varying levels of system uncertainty. Equally important is the evaluation of design features. Such as uncertainty communication, disagreement prompts, and post-error feedback. These can support more accurate alignment between reliance and system capability. Ultimately, designing for calibration means helping users rely on AI in proportion to what it can do safely, rather than simply increasing confidence in its outputs. Achieving this balance requires collaboration among human-factors specialists, domain experts, and AI developers to ensure that decision support systems remain transparent, adaptive, and aligned with human judgment in contexts where accuracy and accountability are critical.

The review study has its limitations by the small number of included studies and by heterogeneity in how trust and reliance are operationalized. Most reviewed experiments rely on framed rather than materially consequential risk, and few employ longitudinal designs. These limitations reflect gaps in the literature rather than shortcomings of individual studies.

AUTHOR ACKNOWLEDGMENT

The authors used automated language tools to improve grammar and clarity. All conceptual development, analysis, and interpretation are the authors' own.

REFERENCES

Choudhury, A. and Shamszare, H. (2023). 'Investigating the impact of user trust on the adoption and use of ChatGPT: Survey analysis', *Journal of Medical Internet Research,* 25, e47184. Available at: https://doi.org/10.2196/47184

de Visser, E.J., Pak, R. and Shaw, T.H. (2018) 'From "automation" to "autonomy": The importance of trust repair in human–machine interaction', *Ergonomics,* 61(10), pp. 1409–1427. Available at: https://doi.org/10.1080/00140139.2018.1457725

de Visser, E.J., Pak, R. & Shaw, T.H. (2020). 'Trust in automation: Empirical considerations for system design', *Human Factors,* 62(2), pp. 260–279. Available at: https://doi.org/10.1177/0018720819888643

Fahnenstich, H., Rieger, T. & Roesler, E. (2024). 'Trusting under risk: Comparing human to AI decision support agents', *Computers in Human Behavior,* 153, Article 108107. Available at: https://doi.org/10.1016/j.chb.2023.108107

Hoff, K.A. & Bashir, M. (2015). 'Trust in automation: Integrating empirical evidence on factors that influence trust', *Human Factors,* 57(3), pp. 407–434. Available at: https://doi.org/10.1177/0018720814547570

Kim, P.H., Dirks, K.T., & Cooper, C.D. (2009). 'The repair of trust: A dynamic bilateral perspective', Academy of Management Review, 34(3), pp. 401–422.

Klingbeil, A., Grützner, C. & Schreck, P. (2024). 'Trust and reliance on AI: An experimental study on the extent and costs of overreliance on AI', *Computers in Human Behavior,* 160, Article 108352. Available at: https://doi.org/10.1016/j.chb.2024.108352

Kuang, X., Vera, A.H., Goldberg, J.H., & Zhu, J. (2020). 'Inside the black box: The effects of explanation style on trust in automated decision systems', *Human–Computer Interaction,* 35(5–6), pp. 495–552. Available at: https://doi.org/10.1080/07370024.2019.1639712

Lee, J.D. & See, K.A. (2004). 'Trust in automation: Designing for appropriate reliance', *Human Factors,* 46(1), pp. 50–80. Available at: https://doi.org/10.1518/hfes.46.1.50.30392

Mayer, R.C., Davis, J.H., & Schoorman, F.D. (1995). 'An integrative model of organizational trust', Academy of Management Review, 20(3), pp. 709–734. https://doi.org/10.5465/amr.1995.9508080335

Murphy, A.H. & Winkler, R.L. (1977). 'Reliability of subjective probability forecasts of precipitation', Journal of the Royal Statistical Society Series C, 26(1), pp. 41–47.

Omrani, N., Rivieccio, G., Fiore, U., Schiavone, F., & Garcia Agreda, S. (2022). 'To trust or not to trust? An assessment of trust in AI-based systems: Concerns, ethics, and contexts, *Technological Forecasting and Social Change,* 181, Article 121763. Available at: https://doi.org/10.1016/j.techfore.2022.121763

Parasuraman, R. & Riley, V. (1997). 'Humans and automation: Use, misuse, disuse, abuse', *Human Factors,* 39(2), pp. 230–253. Available at: https://doi.org/10.1518/001872097778543886

Shin, D. (2021). 'The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI', *International Journal of Human–Computer Studies,* 146, Article 102551. Available at: https://doi.org/10.1016/j.ijhcs.2020.102551

Wang, C., Liu, S., Yang, H., Guo, J., Wu, Y. & Liu, J. (2023). 'Ethical considerations of using ChatGPT in health care', *Journal of Medical Internet Research*, 25, e48009. Available at: https://doi.org/10.2196/48009

Vasconcelos, H., Jörke, M., Müller, H., & Bernstein, A. (2023). 'The effect of explanation fidelity on user trust and reliance in AI-assisted decision-making', *Computers in Human Behavior,* 139, Article 107501. Available at: https://doi.org/10.1016/j.chb.2022.107501