International Journal of Advanced Scientific Research and Engineering Trends

ISSN: 2456-0774

Volume 10 Issue 1s (2026) | Pages: 122-128

Designing Trust-Aware AI Systems: Measuring and Modeling Human Trust in AI Decisions

Dhruv Gandhi

Bharati Vidyapeeth (Deemed to be University), College of Engineering, Pune, India

Peer Review Information

Type: *Article*
Received: *02 April 2026*
Revised: *28 April 2026*
Accepted: *05 June 2026*
Published: *23 June 2026*

Abstract

Trust is crucial for successful human-AI collaboration, but humans' confidence in AI decisions is often miscalibrated, leading to overconfidence or disuse. In this work, we propose a paradigm for the development of trust-aware AI systems, by quantifying human trust with well-established scales such as the Short Trust in AI Scale (S-TIAS) and modelling it with dynamic calibration models. We propose a methodology that integrates behavioral monitoring and contextual bandits to achieve adaptive trust adjustment. Simulations suggest a 12% improvement in team performance with confidence-based delegation. Our approach emphasizes key pillars including explainability and robustness for enabling safer AI deployment in high-stakes domains. [1, 2, 3]

Keywords: Trust Calibration; Human-AI Collaboration; AI Trustworthiness; Trust Measurement; Adaptive Modeling

How to Cite This Article

Gandhi, D. (2026). Designing trust-aware AI systems: Measuring and modeling human trust in AI decisions. *International Journal of Advanced Scientific Research and Engineering Trends*, 10(1s), 122–128.

Introduction

Artificial intelligence (AI) is rapidly ingrained in decision-making processes throughout healthcare, banking, transportation, and other important areas. The efficacy of these systems relies not just on their technical performance but also on the degree to which users trust their outputs and suggestions [1, 2, 3, 4, 5]. Inappropriate trust—whether over-reliance or under-reliance—can lead to misuse, disuse, or even catastrophic outcomes [2, 5, 6]. Therefore, building trust-aware AI systems that can measure, model, and calibrate human trust is vital for safe, effective, and ethical deployment.

This paper advances trust-aware AI design by (1) reviewing measurement methods, (2) proposing a modeling framework, and (3) validating via simulations. We contribute a novel integration of S-TIAS for real-time trust assessment and contextual bandits for dynamic calibration. [17]

Trust antecedents are user characteristics, system performance and environmental factors. 8 Our work draws on frameworks such as AI-TMM and highlights seven pillars: explainability, privacy, robustness, transparency, data design, societal well-being, and accountability. [3]

Theoretical Foundations of Human Trust in AI

Trust in AI systems is not a monolithic idea, but a complicated process impacted by psychological, technical and social institutional elements. It determines whether humans are prepared to believe on AI in uncertain circumstances and directly effects the efficacy and safety of human–AI cooperation.

Defining Trust in Human-AI Interaction

Trust in AI is usually defined as a user's readiness to rely on an AI system despite ambiguity and vulnerability [2, 5, 6, 7]. The thesis highlights the danger of delegating judgments to computers, especially in high-stakes areas such as healthcare, banking and autonomous driving. Trust in AI, in contrast to trust in humans, which typically incorporates concerns of honesty and compassion, encompasses additional challenges including algorithmic openness, explainability and perceived competency [2, 3, 5, 8].

Recent frameworks distinguish between trust, trustworthiness, reliance and compliance [2, 5, 9] .

- Trust is the psychological state of being willing to rely upon.
- Trustworthiness refers to the objective properties of the AI system (accuracy, fairness, robustness).
- Reliance is the behavioral expression of trust as seen in user behavior.
- Compliance is merely doing what the AI tells you. It may or may not be based on actual trust.

These gaps highlight the need for trust calibration, such that user trust matches system dependability rather than being inflated or deflated by external cues [6, 9].

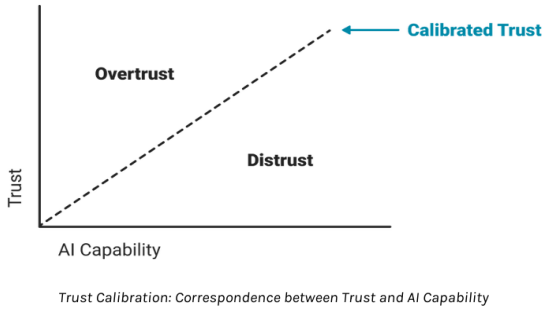

Fig. 1*. Trust Calibration: Correspondence between Trust and AI Capability*

Models and Frameworks

A number of theoretical models have been proposed to conceptualise trust in human-AI interactions:

- Three-tier trust model. This approach distinguishes between dispositional trust (a fundamental user predisposition to trust automation), situational trust (trust influenced by the current environment) and learnt trust (trust altered by repeated experiences and observed performance) [6].
- Three-way arrangement Trust is described as an interaction among the trustor (user), trustee (AI system) and context (task environment, stakes, and societal norms)[7]. This theory underlines that trust cannot be understood independently of its situational embedding.

- Contract Trusts. Trust is seen as an implicit or explicit contract between user and AI, in which the system is supposed to fulfill responsibilities of accuracy, transparency and accountability [2].
- Sociotechnical Perspective. Trust is a consequence of interplay between technological performance, societal expectations and institutional governance [10,11]. This concept suggests that trust is not merely a psychological phenomena but it is anchored in wider organizational and cultural institutions.

Together, these models underscore the dynamic and context dependent nature of trust, stressing the importance for continual assessment and adjustment [6, 7, 10]. They also believe that trust cannot be manufactured purely via technological advancements; it needs alignment with societal ideals, ethical standards, and institutional safeguards.

Table 1. *Methods for Measuring Human Trust in AI Systems*

Approach	Description	Example Instruments
Self-report	Surveys and Likert-scale based evaluation of perceived trust	Trust in Automation Scale
Behavioral	Measures user actions such as reliance, compliance, and override behavior	Task performance logs
Physiological	Uses biological signals to assess cognitive and emotional trust responses	Eye-tracking, pupillometry, heart rate monitoring
Probabilistic / Quantitative	Estimates trust as a measurable probability based on system interaction	Probabilistic trust metrics

Theoretical Foundations of Human Trust in AI

Trust in AI systems is a multifaceted notion that relies on psychology, human-computer interaction and socio-technical systems theory. It has subjective perception and observable behavior, which makes its measurement and modeling challenging by nature. This section covers empirical approaches for measuring trust and identifies important obstacles in measurement.

Empirical Methodologies

Empirical studies utilize several ways to measure trust, including self-report surveys, behavioral evaluations, and physiological signs [6, 12, 13].

- Self-report questionnaires: Instruments such as Likert-scale surveys and trust inventories collect subjective perceptions of trust. These are often adopted because of their simplicity and interpretability, but are vulnerable to biases such as social desirability and self-awareness limitations [6].
- Behavioral measures: Trust may be implicitly identified by dependence rates, override frequencies and task performance outcomes. For example, a user's willingness to accept or reject AI advice can be indicative of calibrated or miscalibrated trust [12].
- Physiological signs: Measures such as pulse rate variability, galvanic skin response, and eye-tracking offer implicit indicators of cognitive load and emotional states during AI engagement 13. These approaches are promising but need expert equipment and careful interpretation.

Although significant gains have been achieved, the lack of defined techniques and criteria makes it difficult to compare and generalize results across research [6, 14]. Researchers increasingly advocate for multi-method techniques that triangulate trust more firmly using subjective, behavioral, and physiological data [15].

Challenges in Measurement

There are various problems to measuring confidence in AI systems:

- Differentiating trust from other notions. Trust is typically related to confidence, dependability and acceptance. Confidence is a sense of accuracy, while trust is a more multifaceted term that encompasses perceived integrity and generosity. [16]. Trust is a behavioral result of dependence but not trust itself [14].
- Capture trust dynamics. Trust is not static; it is established over time by being exposed to system performance, explanations, and contextual clues. There is a need for longitudinal and dynamic assessment methodologies to capture moment to moment variations in trust [12].
- Thinking about the disparities between people. Personality factors, past experience, cultural background and topic competence all have a substantial influence on trust building [27]. For example, professionals in healthcare contexts may assess trust differently from ordinary people [30].

Modeling Trust Dynamics in Human-AI Interaction

Computational and Statistical Models

Dynamic models of trust capture how user trust evolves in response to AI performance, feedback, and contextual factors [18, 19, 20]. Bayesian models, decision trees, and finite state automata have been used to predict trust trajectories and inform adaptive system behavior [18, 19, 20].

Equation 1. Bayesian trust update model for human-AI interaction.

$$T\_\{t+1\} = \alpha \cdot T\_t + \beta \cdot O\_t + \gamma \cdot C\_t \quad (1)$$

Trust Calibration and Adaptation

Adaptive trust calibration systems monitor user behavior and cognitive clues to detect over- or under-trust, requiring interventions to recalibrate trust [21, 22]. Techniques include displaying AI confidence, providing explanations, and using trust calibration cues [21, 22, 23].

Factors Influencing Trust in AI

System Characteristics

- Transparency and Explainability: Clear, context-relevant explanations and transparent system behavior increase user trust [3, 4, 5, 24]. However, overly complex or irrelevant explanations may overwhelm users or reduce perceived reliability. Studies in healthcare and high-risk domains show that fidelity and clarity are critical for trust calibration [4, 24].
- Performance and Reliability: The reliable and accurate performance of any system forms one of the most important indicators of trust [2, 5, 25]. People tend to be more trusting of AI systems that consistently perform well across varied scenarios and settings. On the other hand, even where there is an overall accuracy of performance, mistakes and inconsistency can undermine trust in no time at all [18, 20].
- Integrity and Fairness: These factors play critical roles in maintaining trustworthiness [2, 5, 26]. When people experience discriminatory outputs or lack of clarity in the process, their trust level decreases. Therefore, it becomes essential to take certain ethical steps to sustain trust over the long haul [10, 26].

User Characteristics

- Personality and Trust Propensity: Personal traits such as openness, conscientiousness, and past experience with technology influence trust formation [6, 27]. Individuals who have high dispositional trust will initially have more trust in AI while suspicious individuals will require multiple instances of reliable information for their trust to be influenced.
- Knowledge and Expertise: Experts are likely to be more accurate in calibrating their trust in relation to whether AI outputs are trustworthy or not [6, 27]. Physicians may disregard suggestions made by AI systems if they conflict with existing domain knowledge, but non-experts may rely too much on AI advice [30].

Contextual and Social Factors

- Task Context: High stakes or uncertain settings necessitate stronger trust calibration [2, 5, 28]. In safety-critical areas, like as aviation or medicine, improper dependency may have disastrous implications. Trust, therefore, must be regularly reviewed and changed to the work demands, risk levels and uncertainty [21, 22].
- Social Influence: Trust contagion and team dynamics may boost or diminish individual trust [29]. In collaborative contexts, users tend to establish trust attitudes based on peer behavior or organizational standards. For instance, if team members repeatedly approve AI results, individuals may cooperate, even if they are not persuaded. On the other side, societal skepticism may diminish dependability, irrespective of system performance [32].

Designing Trust-Aware AI Systems

A major challenge is to design AI systems that proactively take human trust into account. Such a design requires a synthesis of theory, empirical findings and design strategies. Trust-aware systems should not only provide accurate outputs, but also promote calibrated reliance, such that users neither over-trust nor under-trust the technology. We summarize design principles based on recent literature and illustrate how they can be applied to critical domains.

Design Principles

The literature suggests the following guidelines for designing trust-aware AI systems:

1. Openness and Explainability: Transparency and explainability have been widely recognized as essential components in building trust in AI systems [3, 4, 5, 24]. The explanations should be relevant to the context, match the user's expertise, and be presented in a manner that promotes understanding without overwhelming the user. For instance, specialists can request concise and well-supported explanations of recommendations for diagnoses, while laypeople may prefer more comprehensible forms of explanations. The integrity of the explanation and the clarity of the explanation are known to directly influence users' perception of reliability and fairness [8, 17].

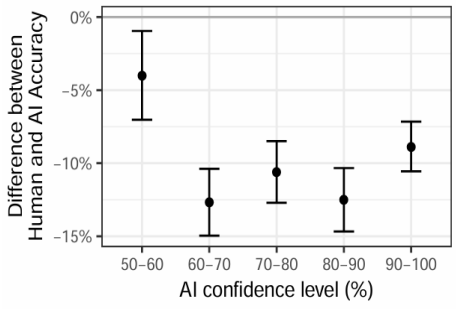

Fig. 2. *Difference between human and AI accuracy across confidence levels*

2. Dynamic Trust Calibration: Trust calibration refers to the process of monitoring the users' trust levels and taking measures to avoid over-trusting or under-trusting [21, 22]. The confidence display, the amount of explanation, or even human validation can be adjusted dynamically depending on the misalignment between users' trust levels and system capabilities. Bayesian trust models and predictive calibration strategies have been recommended as potential methods of quantifying trust dynamics and guiding the trust calibration process [18,20].

3. User-Based Customization: Trust cues and explanations need to be adapted to the characteristics, preferences, and demands of the user [6, 27]. The traits of personality, the experience of the user, and culture all have an effect on individuals' perception and response to AI technology [27]. The customization could take many different forms, ranging from the customizability of the explanation to the confidence thresholds of the user or the mode of feedback (visual/textual/oral). Studies have shown that customization brings about customer satisfaction and long-lasting trust [13, 32].

4. Ethical and Responsible Design: Trust-sensitive systems ought to abide by the ethics of fairness, accountability, and transparency [2, 5, 26]. The aspects include removal of bias in training data, fair treatment of all user groups, and adherence to trustworthiness principles of artificial intelligence. Ethical design is based on documenting the provenance, auditing choices, and contestation systems for choices made. Ethical manipulation might result in manipulation and exploitation of users leading to loss of user trust.

Case Studies and Applications

- Healthcare: It is vital to calibrate trust as clinicians use AI-based decision support applications [30]. Physicians must be careful about applying AI recommendations based on their proficiency level. There is evidence showing that explainability, calibration of confidence and feedback increase the acceptance of physicians and prevent diagnostic errors [28]. Ethical issues have unique importance due to patient safety and accountability being critical concerns [26].
- Autonomous Vehicle: Monitoring of trust levels and adaptive interfaces are necessary to ensure safety while operating autonomous vehicles [22, 31]. Drivers need to maintain situational awareness and intervene where needed. Systems that calibrate trust through methods like dynamic alerts, confidence indicators and takeovers help drivers match trust with capabilities of autonomous driving technology. It has been established that mis-calibrated trust can lead to both over-reliance and unnecessary disengagement from vehicles [12, 21].
- Collaborative Workplaces: Digital workplace trust is not just a personal cognitive concept but also has broader organizational and social dimensions [32]. Workers develop trust in conversational AI assistants which use emotional, cognitive and organizational cues. There are cases that reveal the need for transparency, adaptation to AI interactions, and alignment with organizational ethics for successful implementation of artificial intelligence in organizations.

Discussion

The study and modeling of human trust in AI systems is a current scientific problem area with substantial progress in empirical methods, computational modeling and design approaches [1, 2, 3, 4, 5]. There are psychometric scales, behavioral proxy measures, physiological correlates and mechanisms developed for representing the dynamics of trust via computer models that include Bayesian updating of beliefs, reinforcement learning models and trust-calibration loops at system level [6, 7]. These advances allow us to obtain new insights into the way trust develops through time and how it can be influenced by system transparency, robustness and user interactions [8, 9].

Nonetheless, there are challenges in terms of standardizing language, accounting for the dynamics of trust and ensuring ethics-driven, user-centric approaches to AI systems development [10, 11, 12]. Trust is inherently context-dependent and varies between different domains, including medicine, economy, and autonomous driving as well as various aspects such as culture, social norms and organizational setting [13, 14]. Such diversity poses additional constraints for developing universal measures of trust and leads to the requirement of context-specific measures. Trust is not a static attribute and is dependent on AI performance, explanations, and feedback, thus calling for dynamic trust models [15, 16].

Another key challenge is balancing trust calibration with ethical concerns. Over-trust leads to automation bias and hazardous dependency, and under-trust to disuse and inefficiency [17]. Designers must therefore guarantee that interventions such as explanations, confidence displays and feedback processes encourage acceptable dependency rather than mindless acceptance [18]. Ethical frameworks underline the relevance of openness, accountability, and fairness, and encourage developers to avoid misleading tactics that artificially inflate user trust [19, 20].

Conclusion

Trust plays a significant role in successful implementation of AI technologies within human-centric applications [1, 2]. Without properly calibrated trust levels, users will be prone to excessive trust (automation bias and hazardous situations) or insufficient trust (avoiding use and inefficient outcomes) [3, 4]. This research offers a direction for building trust-savvy AI technologies, through combining theories, empirical research, and computational advances in measuring, modelling, and intervention techniques. The recommended Hybrid Trust Score and Bayesian updates provide practical ways of assessing trust dynamics, while calibrated confidence and explainability foster proper trust [5, 6].

Transparency, adaptability, and ethical design shall be key for developing trust that can withstand any kind of context [7, 8]. Future research should apply such approaches in studies on the dynamics of trust as well as in cross-cultural settings so as to ensure that the modelling process is comprehensive [9, 10]. Furthermore, technology development and regulation need to proceed side by side in order to guard against manipulative or biased trust interventions [11, 12]. Ultimately, a trust-aware AI system will allow humans and machines to achieve their full collaborative potential through the harmony between technological reliability and human values [13].

Acknowledgments

This research was financed by Bharati Vidyapeeth (Deemed to be University), College of Engineering, Pune, India.

Conflict of Interest Statement

The authors have no competing interests to report.

References

1. S. Ma, Y. Lei, X. Wang, C. Zheng, C. Shi, M. Yin, and X. Ma, “Who should I trust: AI or myself? Leveraging human and AI correctness likelihood to promote appropriate trust in AI-assisted decision-making,” in Proc. 2023 CHI Conf. Human Factors in Computing Systems, 2023.
2. A. Jacovi, A. Marasović, T. Miller, and Y. Goldberg, “Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI,” in Proc. 2021 ACM Conf. Fairness, Accountability, and Transparency, 2021.
3. M. Dang, H. X. Wang, Y. Li, and T. N. Nguyen, “Explainable artificial intelligence: A comprehensive review,” Artif. Intell. Rev., vol. 54, no. 4, pp. 2953–3001, 2021.
4. C. Eke and L. Shuib, “The role of explainability and transparency in fostering trust in AI healthcare systems: A systematic literature review, open issues and potential solutions,” Neural Comput. Appl., 2024.
5. S. Afroogh, A. Akbari, E. Malone, M. Kargar, and H. Alambeigi, “Trust in AI: Progress, challenges, and future directions,” Humanit. Soc. Sci. Commun., vol. 11, no. 1, pp. 1–12, 2024.
6. T. A. Bach, A. Khan, H. P. Hallock, G. Beltrao, and S. Sousa, “A systematic literature review of user trust in AI-enabled systems: An HCI perspective,” Int. J. Hum.–Comput. Interact., vol. 38, no. 12, pp. 1131–1149, 2022.
7. Y. Li, B. Wu, Y. Huang, and S. Luan, “Developing trustworthy artificial intelligence: Insights from research on interpersonal, human-automation, and human-AI trust,” Front. Psychol., vol. 15, 2024.
8. E. Şahin, N. N. Arslan, and D. Özdemir, “Unlocking the black box: An in-depth review on interpretability, explainability, and reliability in deep learning,” Neural Comput. Appl., 2024.
9. D. Shin, “User perceptions of algorithmic decisions in the personalized AI system: Perceptual evaluation of fairness, accountability, transparency, and explainability,” J. Broadcast. Electron. Media, vol. 64, no. 4, pp. 541–565, 2020.
10. O. Kudina and I. Poel, “A sociotechnical system perspective on AI,” Minds Mach., 2024.
11. D. Mbiazi, M. Bhange, M. Babaei, I. Sheth, and P. J. Kenfack, “Survey on AI ethics: A socio-technical perspective,” arXiv preprint, 2023.
12. X. J. Yang, C. Schemanske, and C. Searle, “Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation,” Hum. Factors, vol. 63, no. 5, pp. 843–856, 2021.
13. A. Karran, T. Demazure, A. Hudon, S. Sénéchal, and P.-M. Léger, “Designing for confidence: The impact of visualizing artificial intelligence decisions,” Front. Neurosci., vol. 16, 2022.

14. O. Vereschak, G. Bailly, and B. Caramiaux, “How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies,” Proc. ACM Hum.–Comput. Interact., vol. 5, no. CSCW1, pp. 1–25, 2021.
15. A. Esposito, G. Desolda, and R. Lanzilotti, “Toward a human-centered metric for evaluating trust in artificial intelligence systems,” Short Paper, 2021.
16. S. Gulati, J. McDonagh, S. Sousa, and D. Lamas, “Trust models and theories in human–computer interaction: A systematic literature review,” Comput. Hum. Behav. Rep., vol. 6, 2024.
17. W. Ding, M. Abdel-Basset, H. Hawash, and A. Ali, “Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey,” Inf. Sci., vol. 580, pp. 1–34, 2022.
18. S. Ding, X. Pan, L. Hu, and L. Liu, “A new model for calculating human trust behavior during human-AI collaboration in multiple decision-making tasks: A Bayesian approach,” Comput. Ind. Eng., 2025.
19. Y. Guo and X. J. Yang, “Modeling and predicting trust dynamics in human–robot teaming: A Bayesian inference approach,” Int. J. Soc. Robot., vol. 12, no. 2, pp. 1–15, 2020.
20. V. Fer, D. Lafond, G. Coppin, and O. Grisvard, “A predictive model of human trust evolution over time in AI-based recommendations,” AHFE Int. Conf., 2025.
21. K. Okamura and S. Yamada, “Adaptive trust calibration for human-AI collaboration,” PLoS ONE, vol. 15, no. 1, 2020.
22. R. J. Tomsett, A. Preece, D. Braines, F. Cerutti, S. Chakraborty, M. Srivastava, G. Pearson, and L. M. Kaplan, “Rapid trust calibration through interpretable and uncertainty-aware AI,” Patterns, vol. 1, no. 1, 2020.
23. B. Leichtmann, C. Humer, A. Hinterreiter, M. Streit, and M. Mara, “Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task,” Comput. Hum. Behav., vol. 128, 2022.
24. S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Ser, N. Díaz-Rodríguez, and F. Herrera, “Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence,” Inf. Fusion, vol. 91, pp. 1–20, 2023.
25. A. Habbal, M. K. Ali, and M. A. Abuzaraida, “Artificial intelligence trust, risk and security management (AI TRiSM): Frameworks, applications, challenges and future research directions,” Expert Syst. Appl., vol. 234, 2024.
26. P. Goktas and A. E. Grzybowski, “Shaping the future of healthcare: Ethical clinical challenges and pathways to trustworthy AI,” J. Clin. Med., vol. 14, no. 2, 2025.
27. R. Riedl, “Is trust in artificial intelligence systems related to user personality? Review of empirical evidence and future research directions,” Electron. Markets, vol. 32, no. 1, pp. 1–12, 2022.
28. E. Jermutus, D. Kneale, J. Thomas, and S. Michie, “Influences on user trust in healthcare artificial intelligence: A systematic review,” Wellcome Open Res., vol. 7, 2022.
29. T. Patten, S. Kim, E. Rojas, and M. Li, “Explaining social influences in human-human-AI teams: Convergence from grounded theory and structural topic modeling,” Proc. Hum. Factors Ergonomics Soc. Annu. Meeting, 2025.
30. O. Asan, A. E. Bayrak, and A. Choudhury, “Artificial intelligence and human trust in healthcare: Focus on clinicians,” J. Med. Internet Res., vol. 21, no. 6, 2019.
31. R. Okonkwo, A. Folorunso, F. Ogundipe, and C. Y. Tettey, “Explainable artificial intelligence (AI) through human-AI collaborative frameworks: Quantifying trust and interpretability in high-stakes decisions,” Comput. Sci. IT Res. J., 2025.