International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Trust on Trial: How AI Errors Reshape Human Confidence in Technology

Shivam Kumar[1], Shinty P K[2]

[1]CMR University, School of Science and Computer Science, Bengaluru, Karnataka-560043, India
Email: *shivam_kumar[at]cmr.edu.in*

[2]CMR University, School of Science and Computer Science, Bengaluru, Karnataka-560043, India
Email: *shinty.p[at]cmr.edu.in*

Abstract: ***Artificial intelligence (AI) systems increasingly recommend decisions in domains such as healthcare, hiring, and finance, yet these systems are not infallible. How users calibrate their trust and reliance on such systems after witnessing errors—and whether that calibration holds up over repeated exposure—remains only partly understood. This paper proposes a longitudinal, mixed-methods study examining how trust and behavioural reliance evolve across repeated interactions with an AI recommendation system that makes occasional, controlled errors. We review prior work on trust in automation, algorithm aversion, and trust repair, and we outline a research design that varies error timing, error severity, and the presence of explanations across multiple interaction sessions. We propose measuring both self-reported trust and behavioural reliance (e.g., override frequency, response latency) to capture the gap that often exists between what users say and what they do. We discuss expected patterns- including early over-trust, a sharp drop after a first visible error, and partial but incomplete recovery- along with implications for the design of AI-assisted decision support systems.***

Keywords: human-AI interaction, trust calibration, algorithm aversion, AI-assisted decision making, reliance, error recovery

1. Introduction

AI-based recommendation and decision-support systems are now embedded in everyday and high-stakes workflows alike, from suggesting a movie to flagging a suspicious loan application. Their usefulness depends not just on accuracy, but on whether the people working alongside them extend an appropriate degree of trust- enough to benefit from correct suggestions, but not so much that they defer blindly when the system is wrong. This balance, often called trust calibration, is difficult to achieve in practice. Users routinely over-trust systems that appear confident and under-trust systems after a single visible mistake, even when the system's overall accuracy remains high.

Most existing studies of trust in AI examine a single interaction or a short session, giving a snapshot of how one error affects one decision. Far less is known about how trust and reliance behaviour evolve across many interactions, where users accumulate a personal track record of the system's successes and failures. Does an early error leave a lasting scar on reliance even after several subsequent correct recommendations? Does the type of error- a confidently wrong answer versus an uncertain one- change how quickly trust recovers? These questions matter directly for the design of AI systems meant for sustained, everyday use rather than one-off demonstrations.

This paper proposes a study to address that gap. We outline a repeated-interaction experimental design in which participants complete a decision task assisted by an AI recommender across multiple sessions, with errors introduced at controlled points. We track both self-reported trust and behavioral reliance over time, allowing us to observe not only whether trust changes, but whether stated trust and actual reliance drift apart. The contribution of this paper is threefold: (1) a synthesis of the trust-in-automation and algorithm-aversion literature as it applies to repeated exposure; (2) a concrete, replicable methodology for studying trust dynamics longitudinally; and (3) a discussion of the design implications for AI systems that must earn and maintain user trust over time rather than in a single encounter.

2. Related Work

1) Trust in Automation

Trust in automation has long been studied in human factors research, with early work distinguishing trust as a dynamic attitude that responds to a system's perceived reliability, predictability, and competence [1]. This literature established that trust is rarely static: it rises with successful interactions and falls- often sharply- after failures, a pattern sometimes described as being "slow to build, quick to shatter." More recent work has extended these ideas to algorithmic and AI systems specifically, showing that trust judgments are shaped not only by a system's accuracy but by how errors are framed, explained, and timed [2].

2) Algorithm Aversion and Appreciation

A related line of work documents algorithm aversion: the tendency for people to abandon an algorithmic recommender more readily after seeing it err, compared with a human advisor who makes the same mistake, even when the algorithm remains statistically more accurate [3]. Subsequent studies have found this effect is not universal—under certain conditions, such as when users are allowed some control over the algorithm's outputs, aversion is reduced or reversed into algorithm appreciation [4]. These mixed findings suggest that the trajectory of trust after an error is highly sensitive to context, task type, and the interaction design itself, motivating a study that manipulates these factors directly.

3) Effects of Accuracy, Explanation, and Error Type

Work examining model accuracy and trust has shown that users' trust tracks observed accuracy reasonably well within a session, but that this relationship can be distorted by

confidence displays and explanations [5]. Explanations, in particular, have a double-edged effect: they can help users build appropriately calibrated trust, but overly persuasive or complex explanations can also inflate trust beyond what accuracy warrants, especially in clinical and other expert-judgment settings [6]. Few of these studies, however, follow the same users across multiple sessions to see whether these effects persist, strengthen, or fade with repeated exposure- the central question this proposal addresses.

3. Research Questions

Building on this literature, we propose to investigate the following research questions (RQs):

RQ1: How does self-reported trust in an AI recommender change across repeated sessions that include occasional errors, relative to a no-error control condition?

RQ2: Does behavioural reliance (e.g., acceptance rate of recommendations, override frequency, deliberation time) track self-reported trust, or does a gap emerge between what users say and what they do?

RQ3: Does the timing of an error (early vs. later in the interaction history) and its framing (with vs. without an explanation) affect the speed and completeness of trust recovery?

4. Proposed Methodology

1) Study Design

We propose a between-subjects, repeated-measures design with four conditions formed by crossing two factors: error timing (early vs. late) and explanation presence (explanation vs. no explanation), plus a no-error control condition. Each participant completes six sessions over two weeks, spaced to resemble realistic, recurring use rather than a single lab visit.

2) Task and AI System

Participants perform a bounded decision task- for example, estimating whether a short product review is genuine or fabricated- assisted by a recommender that displays a suggested answer and a confidence score before the participant submits their own judgment. The underlying system is a controlled, experimenter-defined "AI" (a Wizard-of-Oz or scripted classifier) so that accuracy and error placement can be manipulated precisely rather than left to a real model's variability.

3) Error Injection

Across the six sessions, the system is scripted to be correct on 85% of trials, matching a plausible real-world accuracy level. In the early-error condition, a salient, high-confidence error is inserted in session 1; in the late-error condition, the same error is inserted in session 5. The explanation manipulation determines whether an error trial is accompanied by a short rationale for the system's (wrong) suggestion.

4) Measures

Self-reported trust is captured after each session using a short, validated trust-in-automation scale. Behavioral reliance is captured continuously through log data: whether participants accept or override each recommendation, how long they deliberate before responding, and how their own accuracy changes when the recommender is available versus (in a final unassisted session) unavailable. Combining these measures lets us examine convergence or divergence between stated trust and actual reliance, directly addressing RQ2.

5) Analysis Plan

We plan to model trust and reliance trajectories using mixed-effects regression, with session number as a time variable and condition as a between-subjects factor, allowing random intercepts and slopes per participant. This approach accommodates the repeated-measures structure and lets us test whether recovery slopes differ significantly by error timing and explanation condition.

5. Expected Findings and Discussion

Based on the reviewed literature, we anticipate several patterns. First, we expect an initial period of over-reliance in all error conditions, consistent with prior single-session findings, followed by a sharp drop in both trust and reliance immediately after the first visible error- larger than the drop predicted by accuracy alone, consistent with algorithm aversion. Second, we expect only partial recovery of behavioural reliance even as self-reported trust rebounds, producing a measurable trust–reliance gap; this would suggest that survey-based trust measures alone may overstate how much people actually rely on a system after it has erred. Third, we expect the late-error condition to produce a smaller relative drop than the early-error condition, since a longer prior track record of correct recommendations may buffer trust against a single mistake- an effect sometimes described as trust inertia. Finally, we expect explanations to speed recovery when they are accurate and specific, but to have little effect, or even a negative effect, when they read as generic justifications for the error.

If confirmed, these patterns would have direct design implications. Systems intended for repeated, long-term use may benefit from front-loading a brief "calibration period" in which early errors are made visible and explained, rather than hidden, so that users build an accurate rather than inflated initial impression of reliability. Interface designers might also track behavioural reliance rather than relying solely on periodic trust surveys, since the two appear to diverge after errors occur.

6. Limitations

The proposed design relies on a scripted rather than a genuinely learning AI system, which sacrifices some ecological validity in exchange for precise control over error timing- a necessary trade-off for isolating the causal effect of error placement. The task domain (review authenticity judgment) is relatively low-stakes; trust dynamics may differ in higher-stakes domains such as medical or financial decisions, where errors carry greater consequence and users may be more risk-averse to begin with. Finally, a two-week window, while longer than most existing studies, still falls short of the months or years-long relationships users may form with deployed AI systems in practice.

7. Conclusion and Future Work

Understanding how trust and reliance evolve over repeated exposure to an imperfect AI system is essential for designing tools people will actually use well over time, not just in a single demo session. This paper has proposed a longitudinal, controlled study to trace that evolution, distinguishing between what users say about their trust and what their behavior reveals about their reliance. Future extensions of this work could examine higher-stakes decision domains, compare scripted systems against real, evolving models, and test specific interface interventions- such as confidence calibration displays or selective error disclosure- designed to shorten the recovery period after a trust-damaging error.

References

[1] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," Human Factors, vol. 46, no. 1, pp. 50–80, 2004.
[2] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," Human Factors, vol. 39, no. 2, pp. 230–253, 1997.
[3] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err," Journal of Experimental Psychology: General, vol. 144, no. 1, pp. 114–126, 2015.
[4] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them," Management Science, vol. 64, no. 3, pp. 1155–1170, 2018.
[5] M. Yin, J. Wortman Vaughan, and H. Wallach, "Understanding the effect of accuracy on trust in machine learning models," in Proc. CHI Conf. Human Factors Comput. Syst., 2019, pp. 1–12.
[6] A. Bussone, S. Stumpf, and D. O'Sullivan, "The role of explanations on trust and reliance in clinical decision support systems," in Proc. Int. Conf. Healthcare Informatics, 2015, pp. 160–169.
[7] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: Challenges and prospects," arXiv preprint arXiv:1812.04608, 2018.
[8] K. Vodrahalli, T. Gerstenberg, and J. Zou, "Uncalibrated models can improve human-AI collaboration," in Proc. Adv. Neural Inf. Process. Syst., 2022.
[9] H. Mehrotra, C. Kettle, B. Vielzeuf, and E. Amid, "Understanding trust dynamics in AI-assisted decision-making: A longitudinal user study," Human-Computer Interaction, 2023.