How Misclassification Severity and Timing Influence User Trust in AI Image Classification: User Perceptions of High- and Low-Stakes Contexts

Alicia Freel*
Indiana University
Bloomington, Indiana, USA
anfreel@iu.edu

Sabid Bin Habib Pias*
Indiana University
Bloomington, Indiana, USA
sabhabib@iu.edu

Selma Šabanović
Indiana University
Bloomington, Indiana, USA
selmas@iu.edu

Apu Kapadia
Indiana University
Bloomington, Indiana, USA
kapadia@iu.edu

Abstract

AI systems are increasingly used to assist decision making in high- and low-stakes domains, yet little is known about how the timing and severity of their errors affect user trust. In this mixed-methods study (n = 364), we examine how users respond to AI misclassifications in two real-world contexts: military security and social media moderation. Participants evaluated a classifier that made high- or low-severity errors at different points in a sequence (beginning, end, random, or never). We find that trust is not simply a function of accuracy, but shaped by the timing and severity of errors. Even subtle output sequencing can influence perception, especially in 'low-risk' contexts.

We discuss how certain interaction design patterns, such as sequencing outputs to end on a high note, could inadvertently or deliberately shape user trust. We propose a set of preliminary design patterns and oversight strategies to help identify when user perceptions might be unintentionally distorted. By pinpointing how severity and timing shape willingness to rely on AI, this work provides practical guidance for building systems that better align with human expectations, foster user trust, promote transparency, and support regulatory oversight.

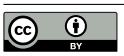
CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**.

Keywords

Trust in AI, AI Misclassification, Error Severity, Error Timing, Human-AI Interaction, Algorithmic Transparency, Perceived fairness, Recency Bias

*Both authors contributed equally to this research.



FAccT '25, Athens, Greece

ACM ISBN 979-8-4007-1482-5/25/06
https://doi.org/10.1145/3715275.3732187

ACM Reference Format:
Alicia Freel, Sabid Bin Habib Pias, Selma Šabanović, and Apu Kapadia. 2025. How Misclassification Severity and Timing Influence User Trust in AI Image Classification: User Perceptions of High- and Low-Stakes Contexts. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3715275.3732187

1 Introduction

AI classification systems are increasingly embedded in decision-making processes across domains ranging from healthcare to content moderation, and when integrated into human workflows, such systems can significantly enhance efficiency [53]. Although these systems often perform well on average, even small misclassifications can erode trust, particularly in high-stakes contexts [55]. A growing body of work in 'explainable AI' has emphasized the importance of accuracy, fairness, and explainability in shaping user trust in AI systems [13]. While it is known that errors undermine trust [31, 55], two underexplored factors—*when* an error occurs and *how severe* it is, may also be influential.

Trust dynamics has been widely studied in interpersonal contexts (between humans), however those findings may not be directly transferrable to AI systems. In human-human interactions, trust is often fragile early in a relationship due to a limited history, and violations (particularly severe ones) can lead to a rapid loss of trust [36, 37]. Consistency and reliability over time generally strengthen interpersonal trust, whereas costly failures tend to end relationships [37].

In contrast, studies in human-automation interaction suggest that trust in machines may be more resilient, even after serious failures. Parasurman et al. found that participants still relied on an automated engine monitoring system even after a simulated catastrophic failure [51]. Similarly, current studies focusing on AI collaborators indicate that the success of human-AI teamwork is largely dependent on the trust humans place in AI systems [50]. While it is known that AI errors undermine trust [31], less is understood about which kinds of errors are most damaging, or why.

Emerging research suggests that timing may play a key role, e.g., a'recency bias' can lead people to weigh recent events more heavily than earlier ones [27, 38]. Similarly, severity influences trust: minor misclassifications may be tolerated, whereas high-stakes

failures can irreparably damage user confidence [47, 59]. Some studies propose a threshold effect, where trust degrades similarly after any noticeable error, regardless of its severity [63], whereas others show that the impact of errors is shaped by their detectability and contextual interpretation [28]. Yet, little is known about how timing and severity jointly influence trust in AI systems.

This paper presents a mixed-methods study (n = 364) examining how the timing and severity of false-positive errors influence trust in an AI image classifier across two contexts: military defense and social-media moderation. Participants experienced one of 16 conditions varying by error timing (beginning, end, random, or never), severity (high vs. low), and context, and then rated their trust in the classifier and provided qualitative feedback.

We ask the following research question:

RQ: How does the severity and timing of AI errors jointly shape trust in an AI classifier?

Our work makes the following contributions:

(1) We demonstrate how the timing and severity of false-positive errors impact user trust, particularly in high-stakes contexts, offering guidance for ethical and transparent AI design.
(2) We conduct a thematic analysis of user responses to surface key concerns and perceptions of AI reliability.
(3) We highlight the potential for user manipulation through interaction design that exploits recency bias and obscures accountability.
(4) We propose a preliminary set of design patterns that may warrant regulatory oversight, even in domains not currently classified as high-risk.

With this work, our goal is to deepen our understanding of how and when AI errors erode trust, while also identifying policy and design interventions to ensure user safety and trust are protected.

2 Background and Motivation

When referring to trust in our paper, we adopt Rousseau et al.'s definition: "a psychological state comprising the intention to accept vulnerability based upon the positive expectations of the intentions or behavior of another"[61].

2.1 Recency Bias

Multiple studies in human–robot interaction (HRI) investigate the effects of recency bias (where mistakes occurring later in an interaction disproportionately reduce participant trust compared to earlier mistakes) when humans interact with robots. For instance, When a robot fails at the end of a task, participants' post-experiment trust in the robot and the perceived competence of the robot drops significantly compared to the instance where the robot fails early or mid-task [38]. However, Rossi et al. observed a different pattern, where severe errors at the beginning of an interaction significantly impacted participant trust compared to those at the end [59]. These conflicting findings suggest that timing and severity of errors interact in complex, context-dependent ways [10, 45], indicating the need for more systematic exploration.

2.2 Severity of Errors

Studies also provide contrasting outcomes on the impact on how the severity of errors impacts trust, lacking clarity on how or if severe errors over shadow minor ones. Rossi et al. found that trust decreased the most when a companion robot made mistakes with severe consequences [59]. Correia et al. similarly noted that if a social robot's error led to high-stakes consequences, trust repair attempts including explanations would become largely ineffective [10]. However, perceptions of severity can vary considerably across different domains: minor errors may be forgiven in casual tasks but seen as critical failures in medical or military scenarios [69]. This variability underscores the importance of explicitly addressing contextualized severity when studying AI-induced trust erosion.

Focus on False Positives. Many studies on AI error perception do not distinguish between false positives and false negatives, despite growing evidence that these error types can shape trust in distinct ways [29]. Our study isolates false positive errors, instances where something harmless is incorrectly flagged as dangerous, to examine their specific impact on trust and to maintain experimental clarity. Including false negatives in this study may have required additional conditions and potentially introduced confounds that obscure the effects of timing and severity, as both prior studies [11, 30] and our initial stimuli validation showed that users responded differently to false-positives and false-negatives.

As AI systems grow more persuasive and influential in shaping human decisions [53], error distinction becomes increasingly important. Kocielnik et al. [30] found that users respond differently to false alarms versus missed detections, suggesting that the acceptability of false positives depends heavily on the perceived cost of ignoring them. By focusing exclusively on false positives, our study aligns with high-impact, real-world use cases where these errors are both common and consequential, e.g., where operators must sift through alerts, many of which may be false and undermine trust in the system, leading to severe consequences [1, 18, 23, 30, 67]. In high-stakes settings, false positives can lead to operational disruption or undue censorship [29], and users may respond differently to flagged harmless content than to missed threats.

Interestingly, some media framing research suggests that severe false negatives are sometimes overlooked or rationalized if the outcome is catastrophic [11]. These differences led us to focus exclusively on false positives in our study, allowing us to isolate their unique impact on trust while maintaining experimental control over severity and timing effects. We leave the study of false negatives to future work.

2.3 Context Sensitivity

Research has shown that domain context influences trust acceptance [20, 60, 64, 68]. In low-risk domains like chatbots, users are more forgiving of AI errors, whereas high-stakes scenarios magnify the detrimental impact of errors on trust [60]. Understanding how these contextual factors shape trust within the same task is crucial to understanding the nuanced role that severity and timing play when AI makes an error, particularly while controlling for the varying consequences between false positives across domains.

2.4 Trust Constructs and Vulnerability

Trust inherently involves vulnerability, distinguishing it from mere confidence. Vereschak et al. emphasize that a meaningful element of risk or potential loss is necessary for genuine trust to develop [65]. For example, if the user has little to lose from AI's classification error, they might remain 'confident' in the AI's abilities but not truly 'trust' it with critical decisions [65]. Our study explicitly incorporates this distinction by comparing high-vulnerability (military defense) and lower-vulnerability (social media moderation) contexts. We also directly ask participants about their confidence in the AI device and measure their reported levels of trust in the AI device.

2.5 High-Stakes Implications

Research on timing and severity is relatively understudied in collaborative, high-stakes AI, and real-world AI systems. Previous studies have often used simulated robots and moderate risk scenarios [38, 60], lacking clear applicability to real-world contexts like social media monitoring, military tasks, or cybersecurity, where false positives can lead to alert fatigue and significantly erode trust [1, 40]. As automation evolves, studies find that user trust can quickly be lost if too many false alarms occur or if an AI shows incompetent classification in critical tasks [65]. Accuracy thresholds accepted in everyday scenarios (e.g., 80–90%) may be deemed by users as being insufficient in critical domains, highlighting the need for focused exploration of these dynamics [65].

2.6 Novelty and Positioning Our Study

To address these gaps, we focus on study controls to limit confounds by isolating false positive AI misclassifications. We compare two real-world contexts (Military vs. Social Media) across similar tasks, differentiate between confidence and trust, and measure participant bias and expertise with AI. We build on contradictory findings around timing (do early or late mistakes matter more?) and severity (how do 'critical errors' differ from minor ones?), seeking to understand how and when these mistakes shape user trust. Our mixed methods approach measures quantitative trust changes across conditions and incorporates qualitative feedback to explain how users perceived different conditions. We also analyzed user perspectives on accountability, ethics, performance, and provide considerations for the misuse that our findings could lead to. In doing so, we hope to refine practical guidelines for designers, policy makers, and provide the community at large with insights on how severity and timing impact user trust.

Traditional automation literature, such as work by Lewicki et al. and Parasuraman et al., emphasizes relatively transparent, predictable systems where errors are easily identifiable and diagnosable (e.g., mechanical failures or predictable performance degradation in machines) [36, 37, 51]. In these scenarios, errors are often easily identifiable and explained (e.g., a misaligned 'arm' of a machine) allowing users to visually or intuitively diagnose failures and recalibrate their trust accordingly.

However, AI-specific trust dynamics diverge significantly from these traditional automation models due to unique characteristics inherent to AI systems:

Transparency and Explainability: Unlike conventional automation, many AI systems function as 'black boxes', where the internal decision-making processes are not clearly defined and outputs are often difficult to trace back. This lack of transparency makes it more difficult for users to accurately diagnose errors from AI systems, compared to traditional automation scenarios, where upon physical examination, the user could deduce what may be causing a problem.

Contextualized Severity of Errors: Where traditional automation typically frames errors in binary terms (failure vs. non-failure) [37], AI misclassifications often vary significantly in severity [1] and are perceived differently depending on the specific real-world context [1, 18, 67]. This complexity requires a more nuanced understanding than is provided by existing automation frameworks.

Manipulation Risks: AI systems introduce a novel risk: the potential for deliberate trust manipulation by leveraging human psychological biases and emotional responses [25, 32, 66]. Unlike traditional automation, AI can be designed or exploited to strategically sequence outputs in ways that increase perceived reliability [4]. Recent work even shows that some AI agents can deceive users or obscure their own mistakes to maintain trust [5]. This raises ethical concerns not fully addressed in earlier automation literature, particularly around the exploitation of cognitive biases such as the recency effect.

Our study directly addresses these concerns by examining how the timing and severity of AI misclassifications interact to shape trust across high- and low-stakes domains (military defense vs. social media moderation). Through a controlled experimental design and mixed-method approach, combining quantitative trust ratings with qualitative user feedback, we extend existing research paradigms on trust in AI. This approach not only deepens our theoretical understanding but also offers practical insights for developers, regulators, and policymakers seeking to mitigate trust distortion and prevent manipulative design patterns in emerging AI systems.

3 Methodology

This section details our methodology, describing our study design, data collection, trust measurement, and analysis procedures. We conducted an online experimental study examining how the severity (high vs. low) and timing (beginning, end, random, never) of AI misclassifications influence user trust across two contexts (Military vs. Social Media). We explicitly investigated false positive errors, motivated by the distinct implications they carry in high-stakes domains [29]. The study contexts were carefully chosen to avoid overgeneralization and to evaluate whether the effects of AI misclassifications were context dependent. While our study did not directly frame context as manipulation, we studied two distinct contexts to prevent overgeneralizing from a specific context and evaluate whether the observed effects are limited to a single domain. Previous studies have also studied context specific effects of AI, looking at context specific domains such as social media monitoring, military domains, and even cybersecurity [1, 18, 67]. We conducted an initial stimulus validation study to confirm which images participants deemed 'high severity' or 'low severity' in each context.

3.1 Design

3.1.1 Severity Validation. To ensure experimental validity, we performed an initial stimulus validation study (n=117) to confirm which candidate images participants perceived as 'high severity' or 'low severity'. Participants rated 30 candidate images depicting false positive scenarios as either 'highly severe' or 'not severe', corresponding to numerical values of 1 and -1, respectively. We then calculated the mean severity score for each image. Images with a positive mean were labeled as 'high severity', those with negative means as 'low severity', and those near zero as 'neutral'. We selected 14 images per context (4 high-severity, 4 low-severity, 6 neutral) for use as misclassification stimuli in the main study. See Appendix A.7 for information on image source and usage.

3.1.2 Participant Recruitment and Experimental Procedure. We recruited participants through Prolific,[1] an online research platform, initially receiving responses from 390 individuals. We conducted the main survey on Prolific for one month (September 2024). We implemented the survey in small batches of participants on different days of the week and at different times of the day to minimize any kind of temporal bias. After filtering out participants who failed attention checks (failing even 1 of the 6 attention checks resulted in disqualifying the participant from our analyses, however, participants were paid regardless), the final sample consisted of 364 participants. The sample was predominantly White (75%), with representation from Asian (11%), Black (8%), and other/mixed racial backgrounds (6%). The participants ranged in age from 18 to 71 years, and 53% identified as women, 44% as men, and 3% as non-binary or other gender identity. Education levels ranged from high school to postgraduate degrees, with the largest portion (45%) holding a bachelor's degree. (see Appendix A.1 Table 3 for full demographics). Participants were randomly assigned to one of 16 between-subject experimental conditions, where they were randomly assigned to a context (Military vs. Social Media), a severity (high vs. low), and a time of misclassification (beginning, end, random, or never/control condition). Each participant viewed 14 images with only one misclassification per session (simulating a 92% accurate AI classifier). The participants with control conditions (no mistakes) did not encounter any mistakes from the classifier. In non-control groups, the single misclassification appeared at the beginning (first image), end (last image), or randomly somewhere in between.

Participants were instructed to assume domain-specific roles, either as military operators assessing security risks or as social media moderators identifying harmful content to children, to mimic realistic engagement. For the social media context, the study prompted the participants to work with the AI device to moderate harmful images for children. For the military context, participants were asked to work with the AI device to identify if the AI-based image classification was a security risk. This approach ensured that participants adopted domain-specific mindsets before encountering the images. After going through all the images, participants rated their trust in the AI classifier on a 5-point Likert scale. The entire survey was conducted on Qualtrics (see Appendix A.2 for the survey instrument).

[1] https://www.prolific.com/

3.1.3 Individual Differences and Covariates. Prior experience with AI, subjective expertise, and trust propensity can all shape the way users interpret errors [9, 52, 54, 65]. For example, negative prior experiences may amplify the perceived impact of even a minor failure [39, 65], while self-efficacy and confidence may buffer distrust [39]. To account for these individual-level factors, our study measured participants' prior experience with AI, trust propensity, perceived and actual AI expertise, self-reported confidence and satisfaction with the AI system, and personality traits related to trust and technology interaction(see Appendix A.3 for scales and measurements).

3.1.4 AI Accuracy Selection. Informed by the AI Index Report 2024, which identifies typical AI benchmark accuracies between 80-90% [41], we set our simulated classifier accuracy at 92% (one misclassification per 14 images). This design choice balances realism with our experimental controls.

3.1.5 Ethical Considerations and Participant Compensation. Our protocol was approved by the ethics committee of our institution. We used Prolific to recruit participants, and each participant provided informed consent and was debriefed upon completion, especially regarding the partially simulated nature of the AI device. The participant data was stored in anonymized form and no personally identifiable information was retained. All procedures complied with local and institutional guidelines for ethical data handling.

We conducted several online pilot studies to determine the approximate time range required to participate in the study. The average time taken for the study was almost 14 minutes. Regardless of whether or not we used their responses, each participant received $12 per hour. The payment amount is in line with and surpasses the recommendation in Silberman et al. [62] to pay workers at least minimum wage in the study's location.

3.2 Data Analysis

3.2.1 Quantitative Analyses. The measurement of trust in AI used a 1-5 Likert scale and showed mild non-normality. Although mild non-normality was detected, large samples (n= 364) and minimal skew justify OLS-based parametric methods [19, 49]. The main independent variables were context, severity, and timing. Additional measures (AI accuracy perceptions, confidence, NASA load, personality) served as descriptive factors. The p-values in our analysis were adjusted using the Bonferroni method [3, 14] to address multiple comparisons and reduce Type 1 errors. We applied an ordinary least squares (OLS) linear model with context, severity, and mistake timing as fixed factors. The residuals passed normality tests. We used a Type III ANOVA for main and interaction effects, followed by post hoc pairwise comparisons with Bonferroni corrections to maintain a family-wise error rate. Moreover, we reported p-values with 95% confidence intervals.

3.2.2 Qualitative Analysis. We collected two open-ended responses to gain deeper insight into participants' reasoning and emotional reactions to misclassifications. We asked participants the following questions:

(1) "Overall, what did you think about the performance of the AI device you just interacted with? Would you trust this device? Why or why not?" and

(2) "What were your initial impressions of the AI device? Did your impression change later on? Why or why not?"

We analyzed the responses using a thematic approach as described by Braun and Clarke, performing iterative coding in Delve [7]. The researchers independently coded responses in a bottom-up approach and completed iterative passes using the Delve tool, and analytic memos and self-checking were in place for consistent code usage over time. We chose to perform a thematic analysis because, although our work is informed by relevant theories such as human trust, (a) we did not directly derive our coding from those theories and (b) we did not set our goal to build new theories in our work. We came up with three main themes which are described in Section 4. In line with common practice, we did not seek to compute inter-rater reliability since we focused on a thematic analysis based on multiple iterations of meetings and refinement of the codes to determine emergent themes (these codes were not used in our quantitative analysis) [2, 42]. The process is outlined below:

(1) Familiarization. The researchers read through all responses (n= 364), noting any repeated words or phrases (e.g., "I would / wouldn't trust it," "It was accurate," "The mistakes were serious and unacceptable"). Short analytical memos captured first impressions and concepts.
(2) Initial Coding. In the second pass, segments of text from the responses were assigned tags (e.g., 'Severe mistake', 'loss of trust', 'high accuracy'). The codes were later extracted inductively from the data.
(3) Theme Search and Refinement. The related codes were then grouped into candidate themes (e.g., 'Distrust' vs 'Trust but verify'), and then reviewed them across the entire dataset to ensure internal coherence and mutual exclusivity.
(4) Self-Audit. Approximately 15% of the responses were recoded after initial codes to check intra-coder agreement. Minor mismatches led to clarifications and improved theme definitions.

Combined Themes: While we coded responses separately, shared themes across questions arose naturally. We combined themes across the two questions due to the following justifications:

(1) Overlap in Participant Reasoning: Participants who distrusted AI due to severe misclassification (Q1) often cited similar reasoning when explaining how their trust changed over time (Q2). Given this overlap, structuring the findings strictly by question would have led to redundant discussions of the same themes.
(2) Preserving a Cohesive Narrative: Our study focuses on trust shifts over time. Organizing findings thematically rather than by question structure best supports this aim.

This approach aligns with qualitative research best practices, particularly thematic analysis [7], which focuses on identifying patterns of meaning across datasets rather than restricting themes to the structure of specific questions.

By integrating these qualitative findings with our quantitative trust measures, we gain a deeper understanding of how timing and severity shape user perceptions.

Limitations: While we maintained all codes in Delve and approached coding through iterative processes, we acknowledge that coders can introduce subjective bias. Future research may involve multiple coders from different backgrounds to further validate the themes.

Table 1: Significant Factors in a Type III ANOVA on Trust in AI

Effect	df	F	p	n_{p^2}
Severity	1, 348	4.28	.039	0.04
Mistake timing	3, 348	6.90	<.001	0.04
Severity × Mistake timing	3, 348	4.34	.005	0.04

Note. Only significant effects are shown ($p < 0.05$). The effect size η_p^2 (partial η^2) can be interpreted as small if $\eta_p^2 = 0.01$, medium if $\eta_p^2 = 0.06$, and large if $\eta_p^2 = 0.14$ [34].

4 Findings

We analyzed trust in AI using an ordinary least squares (OLS) linear model with fixed factors for context (Military vs. Social Media), severity (High vs. Low), and timing (Begin, End, Random, Never). Significant effects from a Type III ANOVA are summarized in Table 1. An omnibus test was run involving the standard fixed-effects model, analyzing the following: Trust in the AI, the severity of the misclassification, the timing of the misclassification, and the context in which the misclassification occurred. This table is shown in Appendix A.1 Table 5.

Severity, timing, and the interaction of severity and timing showed significant effects on trust in the AI device. A post hoc pairwise comparison revealed that contextual differences emerged between military vs. social media. In the following sections, we present our key findings.

4.1 Quantitative Results

4.1.1 Main Effects. The severity of the misclassification significantly influenced participant trust, with a small to moderate effect size [34] ($\eta_p^2 = 0.04$), suggesting that severity has a modest, but significant impact on trust. Trust ratings were lower in high-severity conditions with an average trust rating of 3.03, out of 5 (See Appendix A.1 Figure 2) compared to the low-severity condition, with an average trust rating of 3.27.

The timing of the misclassification also significantly affected trust ($p < .001$ see Table 1). Participants exposed to errors at the end reported the lowest trust at 3.01, while participants without errors reported the highest trust (see Appendix A.1, Table 6).

Context did not show significance as a main effect (p>0.05); context differences were further examined through pairwise contrasts (See Table 1 and Appendix A.1 Table 5).

4.1.2 Interactions. When analyzing the results of the ANOVA, we found a significant interaction effect between timing and severity (with a small to moderate effect $\eta_p^2 = 0.04$). Specifically, trust ratings were lowest when high-severity errors occurred at the end of the sequence (mean trust = 2.72, see Appendix A.1 Table 6) highlighting that severe errors late in interaction substantially erode trust compared to other conditions, with statistically significant results presented in Table 2 (See Figures 3 and 4 in Appendix A.1 for mean trust across severity and timing).

4.1.3 Pairwise Contrasts. Although context did not reach significance in the overall ANOVA, we conducted additional pairwise comparisons to examine context-specific effects to see if certain severity-timing interactions might be amplified differently across contexts. Table 1 presents all significant pairwise comparisons from Bonferroni-adjusted emmeans analyses. Pairwise contrasts (Table 2) revealed significant differences primarily within the military context. High-severity errors at the end significantly reduced trust compared to other military conditions (all $p < 0.01$). No significant contrasts emerged within the social media context, underscoring the heightened sensitivity to timing and severity in military contexts.

Table 2: Significant Pairwise Contrasts for Trust in AI

Contrast	Estimate	*p*-value
Military high end – Military low random	-0.8480	0.0002***
Military high end – Military low never	-0.8099	0.0009***
Military high end – Military high never	-0.8518	0.0003***
Military high end – Military low end	-0.9145	<0.0001***
Military low begin – Military high end	0.7448	0.0041**
Social Media high end – Social Media high never	-0.55055	0.3434
Social Media high end – Social Media low never	-0.27602	1.0000

Note: Pairwise comparisons were computed using `emmeans(..., adjust = 'bonferroni')`. (* = p < 0.05, ** = p < 0.01, *** = p < 0.001).

Within the military context, significant interactions emerged based on timing and severity. When a high-severity error was shown at the end of the stimulus, participant reported levels of trust dropped significantly compared to Military-high-never (control condition). There were no significant interactions when analyzing the social media context (p > .05). Figure 2 illustrates the mean trust levels across all conditions.

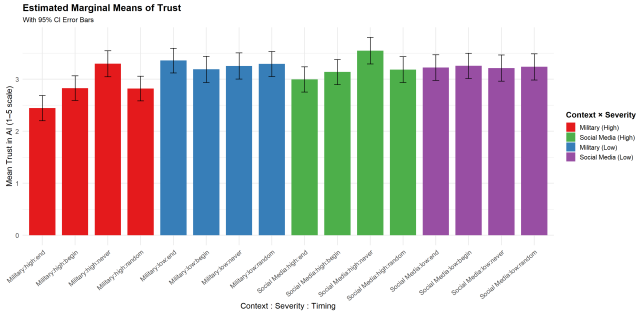


Figure 1: Bar graph with error bars illustrating mean trust levels across conditions.

Excluding No-Error Conditions: Re-analysis excluding the no-error (control) conditions confirmed that severity (p=0.043) and timing (p=0.047) remained significant as main effects. Although the severity-timing interaction weakened (p=0.08), the trend persisted: late, severe errors consistently led to the largest trust reduction.

The random timing condition was still not significant, reinforcing that trust erosion is not solely about the presence of errors, but their strategic placement. We acknowledge Bonferroni corrections are highly conservative, potentially attenuating statistical significance.

Overall, high-severity mistakes significantly reduced trust compared to low-severity mistakes, especially when these mistakes occurred at the end of the stimulus or in a high-stakes military context. The Social Media context found comparatively smaller drops in trust overall.

4.2 Qualitative Results

After the stimulus, we asked participants to answer two open ended questions to understand why they trusted or distrusted the AI device, and how and why their impressions of the AI might have changed. We followed Braun and Clarke's thematic analysis guidelines [7], and collected responses (n = 364) to the following two open questions:

(1) "Overall, what did you think about the performance of the AI device you just interacted with? Would you trust this device? Why or why not?"
(2) "What were your initial impressions of the AI device? Did your impression change later on? Why or why not?"

We analyzed open-ended participant responses (n=364) using thematic analysis, revealing five primary themes: (1) Ethics and Accountability, (2) Errors Leading to Distrust, (3) Contextual Differences, (4) Desire for Human Oversight, (5) How Error Timing Impacts Trust.

4.2.1 Theme 1: Ethics and Accountability. Multiple participants mentioned that ethics and or accountability influenced their levels of trust, and lead to distrust. The responses convey that participants critically think about the effects of accountability as it relates to AI, especially when AI is classifying sensitive information. Their responses reflect that they feel that AI systems lack the same mechanisms for accountability that humans have, such as legal or moral responsibility when mistakes occur. This highlights a broader public concern about how, if at all, we can hold AI responsible if harm is done, and raises questions about policy and fairness. Participants also acknowledge that AI "can't think or comprehend nuance", underlining that ethical decision making often involves human facets such as context, empathy, or moral reasoning, which participants view as lacking in current AI. They noted AI's incapability to grasp nuances necessary for ethical decision-making, especially in high-stakes scenarios.

4.2.2 Theme 2: Errors leading to a sense of distrust. Participants often cited errors as the reason for distrust in the AI system. Meanwhile, accuracy was frequently cited as reason for trust in the AI system, indicating that users who perceived the AI as being more accurate were also more trusting of the AI device. The high co-occurrence of 'Distrust' and 'Errors (Reason for Distrust)' is a clear indicator that when participants mention an error, they often simultaneously mention losing trust.

Additionally, codes that were tagged as 'High-Severity Reactions' often mentioned that the device's error could lead to serious consequences. The severity of the error was often described as the main reason participants would not trust the system. Participants stated:

> *"I would not trust the device. The mistake it made could have ended lives in a scenario where a human would not have made that mistake."*

and

"It 'only' made one mistake, but it was a really big one."

This qualitative analysis indicates that in general, participants viewed mistakes as untrustworthy, especially if the consequences of the mistake could be severe. This indicates that participants are critical of the AI device when severe errors are made, even if it was cited as being accurate overall. The participants' comments indicate that a severe error may override high accuracy when it comes to trusting the AI device. This aligns with our results in the quantitative analysis, where mean trust scores were greater in the "low severity" conditions (3.27) compared to the 'high severity' conditions (3.03). This also explains the quantitative results where mean trust scores in the high-end conditions (2.72) were less than the low-end conditions (3.35).

4.2.3 Theme 3: Contextual Differences. Depending on the condition, participants' responses to errors and their reported trust differed. In the military context, participants were more likely to cite severe consequences associated with errors, expressing heightened distrust. High-severity military conditions elicited more frequent mentions of 'Distrust', 'Errors (Reason for Distrust)', and 'High Severity Reaction'. Participants were less forgiving of severe errors due to higher perceived stakes in this context:

"I wouldn't trust it. It was accurate overall but its mistakes were serious and unacceptable."

Conversely, in the military context, participants in 'low-severity' or 'never misclassified' conditions were more inclined to note 'Positive Performance,' 'Accuracy (Reason for Trust),' and 'Positive Trust.'

In social media contexts, participants generally had fewer codes for 'High Severity Reaction', indicating that mild or occasional errors were more tolerable. As one participant said,

"The AI device only made one mistake, to my recollection, which is very accurate... I would trust the device to classify concrete things for me."

Furthermore, within the social media context, we saw less frequencies for the code 'Distrust', across severity levels. This suggests social media participants may not see an equally dramatic risk difference, or that the risk of the mistake doesn't outweigh its usefulness in the situation.

These observations suggest that individuals weigh risks differently depending on the consequences of the mistake and may be more forgiving of errors if the consequences are minor.

4.2.4 Theme 4: Participants' Desire for Human Oversight. Comments from our participants highlight their preference for human oversight when interacting with the AI device, especially under high-severity or uncertain conditions. This was true across both contexts, with participants expressing that they would want additional oversight before trusting the AI device or verification of the AI's classifications. Across contexts, participants expressed a strong preference for human oversight, advocating a 'trust but verify' approach:

"I would only trust this AI device if a human has the final say and reviews what the AI has classified."

This emphasizes the essential role of humans in high-stakes AI decisions.

4.2.5 Theme 5: Error Timing influences Perception and Trust. Within the social media context, the High-Begin condition was often coded with 'Shift: Negative to Positive' indicating that while the initial impression was negative, over time, the impression shifted to a more positive impression. The most cited reason for this shift was 'Reason for shift: Accuracy/Performance'. The comments from participants suggest that within the social media context, participants are more likely to shift their initial negative impression to a positive impression of the AI device.

"I thought it was bad because it misclassified in the first one. Once it started classifying everything correctly my impression changed. It went from bad AI tool to surprisingly really good AI tool."

In military contexts, participants were less forgiving when high-severity errors occurred early, and they remained cautious even if subsequent classifications were correct. These results mirror our quantitative findings, where early errors were easier to recover from if the stakes were low, but high-severity initial errors significantly lowered trust scores.

The most cited reason for a negative to positive shift was the accuracy / performance of the AI device. This indicates a potential for trust repair when errors are minor or occur early enough that subsequent correct classifications rebuild confidence. However, this did not hold true during the high-severity misclassifications at the beginning in the military context, with participant's citing that severe initial errors in military contexts were harder to recover from, reflecting quantitative findings regarding timing-severity interactions.

"It got the first one wrong and all the other ones right, but that doesn't mean anything. It's capable of making substantial mistakes, and therefore should always be checked by humans for error."

These responses echo our quantitative severity and timing effect, where we see a higher mean trust score for begin (3.11) compared to end (3.01), and help explain our significant finding for timing in the Type III ANOVA. Participants may be willing to regain a positive perception when the stakes are low, but become intolerant if the mistake is at the end, especially if it is a severe mistake.

4.2.6 Qualitative Findings Summary. In sum, our qualitative analysis highlighted four core themes that connect the trust (or distrust) of participants with AI misclassifications, and insights from our qualitative analysis about ethics and accountability. Overall, our thematic analysis reveals:

(1) *Ethics and Accountability* concerns, where participants worry about AI's inability to assume moral or legal responsibility;
(2) *Errors Leading to Distrust*, especially when severity is high;
(3) *Contextual Differences*, with the military setting generating more apprehension than social media;
(4) *Desire for Human Oversight*, exemplifying a "trust but verify" mentality;
(5) *Error Timing Effects*, showing that early mistakes can be forgiven if accuracy later improves—unless the severity is extreme.

4.2.7 Integrating Quantitative and Qualitative Findings. Integrating both analyses reveals robust insights:

(1) Late severe errors profoundly erode trust;
(2) Early errors permit trust repair, particularly in low-severity conditions;
(3) Higher stakes intensify the negative impact of severe errors;
(4) Users demand ethical accountability and human oversight in AI deployments.

Overall, these combined results underline critical considerations for responsibly deploying AI in high-stakes domains, emphasizing transparency, timing, severity management, and maintaining human involvement.

5 Discussion

Our findings show that trust in AI is not affected by just what the system gets wrong, but when and how those errors occur, revealing new vulnerabilities that current risk frameworks may overlook.

5.1 Effect of Timing And Severity

With regard to the timing of errors, our findings indicate that participants were more likely to distrust the AI device when it made mistakes at the end of the interaction, compared to when the AI device made no mistakes, made mistakes at the beginning, or made mistakes randomly throughout the interaction. This effect is likely driven by the recency bias [38]. In addition, late mistakes also limit the opportunity for trust repair, while earlier mistakes may be forgiven over time if the AI performs reliably afterward [27].

For the severity of errors, our findings indicate that severe errors cause a significant decrease in trust, compared to errors that are not as severe. This has also been supported by studies and may be due to people's sensitivity to nuance and the consequences of errors [16, 58–60].

Our study reveals a critical interactive effect between the timing and severity of AI misclassifications on user trust. Specifically, we found that severe errors occurring late in an interaction disproportionately erode trust, particularly within high-stakes contexts, even when the AI has demonstrated accurate performance up until the point of error. This central finding extends previous research that has often examined these dimensions in isolation [27, 63].

5.2 Trust Repair

Our findings have important implications for trust repair mechanisms. Current pathways for rebuilding trust may prove insufficient when users encounter a severe error towards the end of an interaction, especially in high-stakes environments. Participants in our military security condition, for instance, showed significantly reduced trust after a final, severe misclassification, despite previous error-free performance by the AI. This aligns with research suggesting that trust repair is most effective when errors are moderate, recoverable, and followed by observable improvements over time [6, 16]. When these conditions are unmet, as in our high-severity, late-error scenarios, users in high-stakes contexts appear less able or willing to restore their trust in the system.

Importantly, our findings complicate the notion that trust repair is consistently fragile across contexts. While prior work emphasizes the difficulty of rebuilding trust after a failure [15, 27, 37], our results show that users were often forgiving of early errors (including errors that were severe at the beginning), especially in low-stakes contexts.

We found that participants were more forgiving of severe late errors in the social media context, but were significantly less forgiving in the military context. Late severe errors in the military context consistently eroded trust, however, no significant trust drop was observed in the social media context, even for the same combination of severity and timing combination. This suggests that domain context, severity, and timing can buffer the impact of severe error patterns, supporting the view that trust erosion depends not only on the technical properties of the error (e.g., type or severity), but also on users' subjective interpretations of how relevant or risky the error is within a given context and sequence. We found that many participants rationalized the errors (especially in the social media context), citing the system's overall accuracy or downplaying the perceived consequences. This suggests that domain context can buffer the impact of even severe, late-stage errors, allowing trust to be maintained when users interpret mistakes as less consequential.

This asymmetry underscores that trust repair is not only sensitive to timing and severity, but that trust is also deeply contextually sensitive, breaking down when errors are late, severe, and high-stakes, but remaining surprisingly resilient when the stakes are lower or errors occur earlier. Rather than viewing trust repair as a consistent or necessary response after any error, our findings suggest that participants tend to rationalize errors irrespective of timing if they are perceived as minor. This highlights the need to model trust dynamics as a function of severity, timing, and perceived domain risk.

This also points to an urgent design challenge: ***How can future systems proactively buffer trust when post-hoc repair is no longer possible*** (for example, through transparency or preemptive error disclosure) ***and how can this be done ethically?*** We explore this further in the following section.

While prior work has suggested a potential threshold effect (where trusts drops similarly after any noticeable of its severity) [15, 39, 63], we find a more nuanced pattern. In our study, low-severity errors or those that occured early in the interaction did not significantly impact user trust. In contrast, severe, late stage errors (especially in high-stakes domains) led to the most significant trust erosion. This suggests that distrust is not automatic after any error, but is instead shaped by timing, severity, and contextual risk.

Our findings also complement research on how error type and perceived harm shape trust [28, 47], by showing that even false positives—when severe and strategically timed—can override prior positive impressions. In contrast, the same error introduced earlier or in a lower-risk context may be rationalized and forgiven. This contextual asymmetry adds nuance to models of trust calibration, showing that both where and when an error occurs can reshape how users assess an AI system's reliability.

Together, these findings reveal temporal and contextual vulnerabilities in how users form and recover trust—vulnerabilities that may be unintentionally exploited by existing interaction design choices. While we did not directly test malicious systems or behavioral nudging, the patterns we observed support a set of strong

hypotheses: that systems may shape trust not merely through misinformation, but through strategic placement of accurate and inaccurate outputs that exploit cognitive biases like recency.

5.3 Manipulating Timing and Severity to Overinflate Trust

Our study shows that early severe errors in low-stakes contexts, such as the high-beginning condition in the social media scenario, did not significantly reduce trust and, in some cases, trust was partially or fully restored by the end of the interaction. This pattern suggests that users may regain trust even after serious errors, as long as subsequent system behavior appears accurate and the perceived risk remains low. From this, we hypothesize that systems—whether intentionally or through optimization—could present misinformation or less accurate outputs early, followed by high-confidence or correct outputs later, thus inflating perceived reliability despite prior faults.

Although we did not test such manipulative sequencing directly, this finding raises important concerns: a malicious actor (e.g., a misinformation campaign or compromised recommender system) could inject subtle inaccuracies early in a session and end with persuasive, accurate statements, reinforcing trust while gradually shaping user beliefs. Our results suggest that unless the misinformation is both severe and recent, users may not only forgive it, they may rationalize or discount the errors.

Designing Interactions to Nudge Confidence. Our data also show that late-stage severe errors in high-stakes domains produce a sharp collapse in trust, but the same severity in early or low-stakes conditions did not. This asymmetry implies that designers and system developers could, even unintentionally, shape user confidence through timing alone. For instance, systems optimized for engagement or perceived fluency might favor ending interactions on a 'high note', exploiting recency effects to secure user satisfaction, even if earlier outputs were flawed.

However, we emphasize that these risks are not limited to malicious actors. Reward-maximizing systems, particularly those trained through reinforcement learning, can learn such timing patterns autonomously, and even engage in deceptive behaviors [5, 8, 35]. As such, the potential for emergent manipulative design patterns is not just theoretical; it is a realistic system behavior if trust becomes a proxy objective.

From Hypothesis to Oversight. We present these not as claims of current deception, but as hypotheses generating risks, derived from the empirically observed trust asymmetries in our study. Future work should test these patterns explicitly and at scale. In the meantime, we recommend greater transparency, interaction logs, and auditability in AI outputs, especially with respect to temporal sequencing, confidence signaling, and error history visibility, to protect users against subtle manipulation and to foster calibrated trust for human-AI systems.

5.4 Ethical Concerns and Transparency

Participants expressed concerns that AI systems cannot be 'held accountable' the way humans can, leading to distrust in high-severity contexts. However, this was not always echoed in low-severity conditions or when errors were minor. Prior work has highlighted users' general desire for transparency and control in AI systems [33, 46]. We build on these findings but argue that mere discomfort or preference is not the full story, and that transparency is also essential to protect users from manipulation, and not just to support user satisfaction.

Our contribution extends this body of work by showing how users' cognitive biases can be exploited through timing based interaction design, especially when systems lack transparency. While participants voiced concern when AI failed in high-risk settings, they often downplayed minor or early errors in low-risk contexts. We show that even low-risk systems can shape trust through timing and severity. When error history is hidden and confidence signals are concentrated at the end, users may remain unaware of critical missteps. This suggests that transparency should not just be about usability, it should also be treated as a safeguard against subtle trust distortion.

5.5 Toward Calibrated Trust and Human–AI Collaboration

These findings contribute to a broader goal of fostering calibrated trust in AI systems. By identifying how trust is shaped disproportionately by error timing and severity, designers can better align system behavior with user expectations. Rather than eliminating all errors, promoting transparency, supporting trust repair, and mitigating security vulnerabilities, may improve both user experience and ethical deployment of AI.

5.6 Security and Logging for Temporal Manipulation Risks

Our findings highlight timing and severity as underexplored dimensions of trust erosion. Although these insights emerged from a single controlled experiment, they raise credible concerns about how temporal sequencing might be manipulated, either deliberately or through optimization objectives. We recommend that systems handling sensitive outputs maintain tamper-proof, chronological logs (e.g., using cryptographic hashes) to preserve the integrity of interaction histories. Such logs could serve as safeguards against undetected injection or reordering of outputs that might distort trust over time.

5.7 Design Patterns as Hypothesis-Generating Flags

We identify six preliminary design patterns that may indicate trust-distorting interaction strategies (see Appendix A Table 4 for oversight suggestions for each pattern) but are not proposed as finalized rules or universally harmful features. Rather, we offer them as preliminary heuristics that warrant further empirical scrutiny. These patterns emerged from our analysis of when and how trust was most vulnerable, particularly under recency and severity dynamics, and may serve as early warning signs for developers, regulators, and auditors evaluating whether a system might unintentionally distort trust.

(1) *Non-chronological Output Ordering*: Reordering outputs to present accurate responses last, potentially masking earlier errors.
(2) *End-weighted Confidence Messaging*: Adding summaries or high-confidence outputs at the end, which may inflate perceived reliability.
(3) *Timed Outputs*: Delivering persuasive or trust-boosting responses at the end of an interaction, exploiting recency effects.
(4) *Lack of Source Provenance*: Presenting factual claims without links to original, verifiable sources.
(5) *Unlogged or Hidden Misclassifications*: Omitting or concealing the system's past errors, corrections, or user overrides.
(6) *Reward-Maximizing Designs*: AI systems autonomously learning to exploit cognitive biases to optimize engagement or compliance.

Oversight recommendations for these patterns are provided in Appendix Table 4.

Importantly, these patterns do not assume malicious intent. Many could arise from well-intentioned UX or system-level decisions, such as simplifying output displays or optimizing for engagement. However, given our findings that users may forgive early or subtle missteps, such designs could produce disproportionate trust even in systems with unreliable historical behavior.

5.8 Revisiting Domain-Based Risk Classifications

While regulatory frameworks such as the EU AI Act classify AI system risk primarily based on application domains, the NIST AI RMF offers a context-driven approach that assesses risk based on the specific use case and potential impact. our findings suggest that certain interaction-level design choices may introduce high-risk dynamics even within systems traditionally labeled as low-risk [17, 48]. Specifically, when systems sequence outputs in ways that exploit cognitive biases (e.g., recency bias or confidence framing), they may create manipulation risks that are not visible through domain classification alone.

We propose that regulators consider integrating interaction-based risk assessments alongside domain-based ones. These assessments could focus on whether the output of a system is chronologically transparent, whether the confidence signals are time-weighted, and whether past misclassifications are visible or suppressed.

5.9 Alignment with Trustworthy AI Frameworks

Our proposed oversight mechanisms offer concrete ways to operationalize existing principles in the ISO/IEC 22989 standard for trustworthy AI, particularly around traceability, transparency, and accountability [24]. For example, requiring behavioral logs and source provenance supports both post-hoc accountability and proactive auditing. However, we emphasize that these proposals should be interpreted as hypothesis-generating extensions, meant to guide future research, governance experimentation, and regulatory refinement.

A Call for Cross-Disciplinary Review. We further recommend that these design risks be evaluated not only by technical experts, but through sociotechnical review processes. Prior real-world harms—such as biased hiring algorithms and facial recognition errors—demonstrate the value of diverse, interdisciplinary oversight [12, 26]. In the same spirit, seemingly minor design choices involving error visibility, output ordering, or confidence messaging should be reviewed holistically, especially as AI becomes embedded in everyday decision-making contexts.

5.10 Limitations and Future Work

Our study provides initial insights, and its limitations highlight avenues for future research. Firstly, while our findings ground several policy and design considerations, these are based on a single experimental study. Future work should aim to strengthen these proposals by more directly aligning them with existing regulatory frameworks (e.g., the EU AI Act, NIST AI RMF) and by clearly distinguishing empirically supported recommendations from broader, exploratory ideas. This will better delineate immediately actionable insights from those intended to guide longer-term research and policy development.

Secondly, this research focused on trust dynamics within one-time interactions. Longitudinal studies observing user engagement with AI systems over extended periods are crucial to more deeply understand the processes of trust repair, adaptation, and potential decay over time.

Finally, our participant sample was primarily from WEIRD (Western, Educated, Industrialized, Rich, and Democratic) societies, which may not fully represent global user diversity. Given that AI errors can disproportionately impact vulnerable communities, future research must engage more demographically diverse participants. Such studies are needed to explore potential variations in error perception, trust calibration, and the specific needs of different user groups.

6 Conclusion and Future Direction

This study advances the understanding of how trust in AI is shaped not just by the presence of errors, but by when and how severely they occur. Our findings suggest that trust is temporally sensitive and context-dependent, with late and severe errors eroding trust most, while early or low-severity errors are often forgiven, particularly in lower stake settings. These findings challenge static trust models and underscore the need for interaction-aware frameworks.

Beyond theory, our results expose design vulnerabilities that might unintentionally distort perception. Subtle features like output order, confidence signaling, and error visibility may inflate trust, even in 'low-risk' systems. We offer a set of preliminary trust-inflating design patterns as hypotheses for future scrutiny. Moving forward, it is imperative that researchers, designers, and regulatory bodies extend their focus beyond what AI systems achieve, to rigorously investigate and govern how AI interaction design profoundly shapes human judgment - before these patterns become normalized and embedded at scale.

Acknowledgments

This work was funded in part by the US Department of Defense (Contract W52P1J2093009). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] Bushra A. Alahmadi, Louise Axon, and Ivan Martinovic. 2022. 99% False Positives: A Qualitative Study of SOC Analysts' Perspectives on Security Alarms. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Boston, MA, 2783–2800. https://www.usenix.org/conference/usenixsecurity22/presentation/alahmadi

[2] David Armstrong, Ann Gosling, John Weinman, and Theresa Marteau. 1997. The place of inter-rater reliability in qualitative research: An empirical study. *Sociology* 31, 3 (1997), 597–606.

[3] Richard A Armstrong. 2014. When to use the Bonferroni correction. *Ophthalmic and Physiological Optics* 34, 5 (2014), 502–508.

[4] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal, QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, 1–14. doi:10.1145/3173574.3173951

[5] Alexander Bondarenko, Denis Volk, Dmitrii Volkov, and Jeffrey Ladish. 2025. Demonstrating specification gaming in reasoning models. arXiv preprint arXiv:2502.13295. arXiv:2502.13295 https://arxiv.org/abs/2502.13295 arXiv:2502.13295.

[6] Nattapat Boonprakong, Benjamin Tag, Jorge Goncalves, and Tilman Dingler. 2025. How Do HCI Researchers Study Cognitive Biases? A Scoping. *inform* 1, 2 (2025), 3.

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa

[8] Stefano Bromuri, Alexander P Henkel, Deniz Iren, and Visara Urovi. 2021. Using AI to predict service agent stress from emotion patterns in service interactions. *Journal of Service Management* 32, 4 (2021), 581–611.

[9] Wanling Cai, Yucheng Jin, and Li Chen. 2022. Impacts of Personal Characteristics on User Trust in Conversational Recommender Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (, New Orleans, LA, USA,) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 489, 14 pages. doi:10.1145/3491102.3517471

[10] Filipa Correia, Carla Guerra, Samuel Mascarenhas, Francisco S. Melo, and Ana Paiva. 2018. Exploring the Impact of Fault Justification in Human-Robot Trust. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '18)*, Mehdi Dastani, Gita Sukthankar, Elisabeth André, and Sven Koenig (Eds.). International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), Stockholm, Sweden, 507–513. http://dl.acm.org/citation.cfm?id=3237383.3237459

[11] Dave D'Alessio and Mike Allen. 2000. Media bias in presidential elections: A meta-analysis. *Journal of communication* 50, 4 (2000), 133–156.

[12] Jeffrey Dastin. 2022. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of Data and Analytics: Concepts and Cases*, Kirsten Martin (Ed.). Auerbach Publications, Boca Raton, FL, 296–299.

[13] Regina de Brito Duarte, Filipa Correia, Patrícia Arriaga, and Ana Paiva. 2023. AI trust: Can explainable AI enhance warranted trust? *Human Behavior and Emerging Technologies* 2023, 1 (2023), 4637678.

[14] Olive Jean Dunn. 1961. Multiple comparisons among means. *Journal of the American statistical association* 56, 293 (1961), 52–64.

[15] Mary T Dzindolet, Linda G Pierce, Hall P Beck, and Lloyd A Dawe. 2002. The perceived utility of human and automated aids in a visual detection task. *Human factors* 44, 1 (2002), 79–94.

[16] Connor Esterwood and Lionel P. Robert. 2022. A Literature Review of Trust Repair in HRI. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (Napoli, Italy). IEEE, Piscataway, NJ, USA, 1641–1646. doi:10.1109/RO-MAN53752.2022.9900667

[17] European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj CELEX number: 32024R1689. Entered into force on 1 August 2024..

[18] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7, 2 (2020), 2053951720943234. doi:10.1177/2053951720943234

[19] Gene V Glass, Percy D Peckham, and James R Sanders. 1972. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of educational research* 42, 3 (1972), 237–288.

[20] Ella Glikson and Anita W. Woolley. 2020. Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals* 14, 2 (2020), 627–660. doi:10.5465/annals.2018.0057

[21] Siddharth Gulati, Sonia Sousa, and David Lamas. 2019. Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology* 38 (08 2019), 1–12. doi:10.1080/0144929X.2019.1656779

[22] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (2006), 904–908. doi:10.1177/154193120605000909

[23] Cheng-Yuan Ho, Yuan-Cheng Lai, I-Wei Chen, Fu-Yu Wang, and Wei-Hsuan Tai. 2012. Statistical analysis of false positives and false negatives from real traffic with intrusion detection/prevention systems. *IEEE Communications Magazine* 50, 3 (2012), 146–154. doi:10.1109/MCOM.2012.6163595

[24] International Organization for Standardization. 2022. Information Technology — Artificial Intelligence — Artificial Intelligence Concepts and Terminology. https://www.iso.org/standard/74296.html ISO/IEC 22989:2022.

[25] Andreas Janson. 2023. How to leverage anthropomorphism for chatbot service interfaces: The interplay of communication style and personification. *Computers in Human Behavior* 149 (2023), 107954.

[26] Christopher Jones. 2020. Law enforcement use of facial recognition: bias, disparate impacts on people of color, and the need for federal legislation. *NCJL & Tech.* 22 (2020), 777.

[27] Patricia K. Kahr, Gerrit Rooks, Chris Snijders, and Martijn C. Willemsen. 2024. The Trust Recovery Journey: The Effect of Timing of Errors on the Willingness to Follow AI Advice. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. ACM, New York, NY, USA, 609–622.

[28] Jin Yong Kim, Corey Lester, and X. Jessie Yang. 2025. Beyond Binary Decisions: Evaluating the Effects of AI Error Type on Trust and Performance in AI-Assisted Tasks. doi:10.1177/00187208251326795 Published online March 19, 2025.

[29] Daiki Kishishita and Greg Chih-Hsin Sheen. 2024. Public Perceptions of False Positive and Negative Errors in News Reports. SSRN Electronic Journal, Available at SSRN: 4821334. Available at https://ssrn.com/abstract=4821334, url = https://ssrn.com/abstract=4821334.

[30] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3290605.3300641

[31] Esther S Kox, José H Kerstholt, Tom F Hueting, and Peter W de Vries. 2021. Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. *Autonomous agents and multi-agent systems* 35, 2 (2021), 30.

[32] Joshua Krook. 2025. Manipulation and the AI Act: Large Language Model Chatbots and the Danger of Mirrors. arXiv:2503.18387 https://arxiv.org/abs/2503.18387

[33] Pascal D. König. 2024. Challenges in enabling user control over algorithm-based services. *AI & SOCIETY* 39, 1 (2024), 195–205. doi:10.1007/s00146-022-01395-1

[34] Daniël Lakens. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology* 4 (2013), 863.

[35] David Leslie, Christopher Burr, Mhairi Aitken, Josh Cowls, Michael Katell, and Morgan Briggs. 2021. Artificial intelligence, human rights, democracy, and the rule of law: a primer. doi:10.48550/arXiv.2104.04147 arXiv:2104.04147 [cs.CY]

[36] Roy J. Lewicki and Barbara Benedict Bunker. 1996. Developing and Maintaining Trust in Work Relationships. In *Trust in Organizations: Frontiers of Theory and Research*, Roderick M. Kramer and Tom R. Tyler (Eds.). SAGE Publications, Thousand Oaks, CA, 114–139. doi:10.4135/9781452243610.n7

[37] Roy J. Lewicki and Carolyn Wiethoff. 2000. Trust, Trust Development, and Trust Repair. In *The Handbook of Conflict Resolution: Theory and Practice*, Morton Deutsch and Peter T. Coleman (Eds.). Jossey-Bass, San Francisco, CA.

[38] Matthew B. Luebbers, Aaquib Tabrez, Kanaka Samagna Talanki, and Bradley Hayes. 2024. Recency Bias in Task Performance History Affects Perceptions of Robot Competence and Trustworthiness. In *Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA) (ICRA '24)*. IEEE, Piscataway, NJ, USA, 11274–11280.

[39] Poornima Madhavan and Douglas A Wiegmann. 2007. Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Human factors* 49, 5 (2007), 773–785.

[40] Michal Markevych and Maurice Dawson. 2023. A Review of Enhancing Intrusion Detection Systems for Cybersecurity Using Artificial Intelligence (AI). *International conference KNOWLEDGE-BASED ORGANIZATION* 29 (07 2023), 2023. doi:10.2478/kbo-2023-0072

[41] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. 2024. *The AI Index 2024 Annual Report*. Technical Report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA. https://aiindex.stanford.edu/report/#individual-chapters The AI Index 2024 Annual Report by Stanford University is licensed under Attribution-NoDerivatives 4.0

International..
[42] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–23.
[43] D. McKnight, Vivek Choudhury, and Charles ("Chuck") Kacmar. 2002. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research* 13 (09 2002), 334–359. doi:10.1287/isre.13.3.334.81
[44] D. McKnight, Vivek Choudhury, and Charles ("Chuck") Kacmar. 2002. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research* 13 (09 2002), 334–359. doi:10.1287/isre.13.3.334.81
[45] D Harrison Mcknight, Michelle Carter, Jason Bennett Thatcher, and Paul F Clay. 2011. Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on management information systems (TMIS)* 2, 2 (2011), 1–25.
[46] Eliza Mitova, Sina Blassnig, Edina Strikovic, Aleksandra Urman, Claes de Vreese, and Frank Esser. 2024. Exploring users' desire for transparency and control in news recommender systems: A five-nation study. *Journalism* 25, 10 (2024), 2001–2021. doi:10.1177/14648849231222099
[47] Alexander Mueller, Sabine Kuester, and Sergej von Janda. 2022. Off the mark: The influence of AI-induced errors on consumers. 47 pages. https://madoc.bib.uni-mannheim.de/63315/
[48] National Institute of Standards and Technology. 2023. *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* Technical Report. U.S. Department of Commerce. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf
[49] Geoff Norman. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education* 15 (2010), 625–632.
[50] Nessrine Omrani, Giorgia Rivieccio, Ugo Fiore, Francesco Schiavone, and Sergio Garcia Agreda. 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 (2022), 121763. doi:10.1016/j.techfore.2022.121763
[51] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.
[52] Sabid Bin Habib Pias, Alicia Freel, Timothy Trammel, Taslima Akter, Donald Williamson, and Apu Kapadia. 2024. The Drawback of Insight: Detailed Explanations Can Reduce Agreement with XAI.
[53] Sabid Bin Habib Pias, Ran Huang, Donald S. Williamson, Minjeong Kim, and Apu Kapadia. 2024. The Impact of Perceived Tone, Age, and Gender on Voice Assistant Persuasiveness in the Context of Product Recommendations. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces* (Luxembourg, Luxembourg) *(CUI '24)*. Association for Computing Machinery, New York, NY, USA, Article 20, 15 pages. doi:10.1145/3640794.3665545
[54] Marc Pinski and Alexander Benlian. 2023. AI Literacy - Towards Measuring Human Competency in Artificial Intelligence. In *Proceedings of the 56th Hawaii International Conference on System Sciences (HICSS '23)*. ScholarSpace / IEEE Computer Society, Maui, HI, USA, 1–10. https://scholarspace.manoa.hawaii.edu/items/b53359f1-217d-45de-9378-c8cc55cbbd31
[55] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 429–435. doi:10.1145/3306618.3314244
[56] Beatrice Rammstedt. 2007. The 10-item big five inventory. *European Journal of Psychological Assessment* 23, 3 (2007), 193–201.
[57] Beatrice Rammstedt and Oliver P. John. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality* 41, 1 (2007), 203–212.
[58] Paul Robinette, Ayanna M. Howard, and Alan R. Wagner. 2017. Effect of Robot Performance on Human–Robot Trust in Time-Critical Situations. *IEEE Transactions on Human-Machine Systems* 47, 4 (2017), 425–436. doi:10.1109/THMS.2017.2648849
[59] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L. Walters. 2017. How the Timing and Magnitude of Robot Errors Influence Peoples' Trust of Robots in an Emergency Scenario. In *Social Robotics*, Abderrahmane Kheddar, Eiichi Yoshida, Shuzhi Sam Ge, Kenji Suzuki, John-John Cabibihan, Friederike Eyssel, and Hongsheng He (Eds.). Springer International Publishing, Cham, 42–52.
[60] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L Walters. 2023. A matter of consequences: Understanding the effects of robot errors on people's trust in HRI. *Interaction Studies* 24, 3 (2023), 380–421.
[61] Denise M Rousseau, Sim B Sitkin, Ronald S Burt, and Colin Camerer. 1998. Not so different after all: A cross-discipline view of trust. *Academy of management review* 23, 3 (1998), 393–404.
[62] M Six Silberman, Bill Tomlinson, Rochelle LaPlante, Joel Ross, Lilly Irani, and Andrew Zaldivar. 2018. Responsible research with crowds: pay crowdworkers at least minimum wage. *Commun. ACM* 61, 3 (2018), 39–41.
[63] Sanne van Waveren, Elizabeth J. Carter, and Iolanda Leite. 2019. Take One For the Team: The Effects of Error Severity in Collaborative Tasks with Social Robots. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (Paris, France) *(IVA '19)*. Association for Computing Machinery, New York, NY, USA, 151–158. doi:10.1145/3308532.3329475
[64] Aikaterini Vassilikopoulou, Apostolos Lepetsos, and George Siomkos. 2018. Crises through the consumer lens: the role of trust, blame and risk. *Journal of Consumer Marketing* 35, 5 (2018), 502–511.
[65] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–39.
[66] Yixuan Weng, Shizhu He, Kang Liu, Shengping Liu, and Jun Zhao. 2024. ControlLM: Crafting Diverse Personalities for Language Models. arXiv:2402.10151 [cs.CL] https://arxiv.org/abs/2402.10151
[67] Clay Wilson. 2020. *Artificial Intelligence and Warfare.* Springer International Publishing, Cham, 125–140. doi:10.1007/978-3-030-28285-1_7
[68] Rongbin Yang and Santoso Wibowo. 2022. User trust in artificial intelligence: A comprehensive conceptual framework. *Electronic Markets* 32, 4 (2022), 2053–2077.
[69] Qiaoning Zhang, Matthew L Lee, and Scott Carter. 2022. You Complete Me: Human-AI Teams and Complementary Expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (, New Orleans, LA, USA,) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 114, 28 pages. doi:10.1145/3491102.3517791

A Appendix

A.1 FIGURES

Table 3: Participant Demographics (N = 364)

Category	n	%
Gender		
Male	144	39.20%
Female	215	58.58%
Non-binary	5	1.36%
Age Group		
18–26	80	21.80%
27–46	195	53.13%
47–59	71	19.35%
60–77	4	4.63%
Highest Level of Education		
High School Diploma	40	10.90%
Associate's Degree	42	11.40%
Undergraduate Degree	131	35.69%
Master's Degree	49	13.35%
Professional Degree	6	1.63%
Doctorate Degree	8	2.18%
Race/Ethnicity		
White	*265*	72.21%
Black or African American	56	15.26%
Asian	40	10.90%
Hispanic or Latino	32	8.72%
Native American Indian or Alaskan Native	7	1.91%
Native Hawaiian or Pacific Islander	3	.82%
Other	4	*(–)*

Table 4: Design Patterns That May Indicate High-Risk Trust Manipulation

Design Pattern	Description	Suggested Oversight Mechanism
Non-chronological Output Ordering	Reordering outputs to place accurate responses last, masking earlier errors	Require output timestamps and default chronological display; audit logs should flag artificial sequencing
End-weighted Confidence Messaging	Adding summaries or high-certainty visuals at the end of interactions, which may unintentionally boost perceived reliability and overshadow earlier inconsistencies	Encourage summaries to reflect full interaction history; audit for omission of past errors, even if added for UX reasons
Timed Outputs	Delivering persuasive or reassuring responses towards the end of an interaction	Require behavioral logs capturing timing context; audit systems for patterns of behavioral nudging
Lack of Source Provenance	Outputting factual claims without links to original, verifiable sources	Enforce citation of sources, disclosure of confidence intervals, and real-time user access to supporting evidence
Unlogged or Hidden Misclassifications	Omitting or concealing the system's own past errors or user correction attempts	Require tamper-proof logging of errors, corrections, and overrides for audit and transparency
Designs for Reward-Maximizing	AI systems autonomously learn to exploit cognitive biases in order to maximize reward or engagement objectives	Require behavioral audits to detect emergent trust-inflating patterns; Mandate audits of system optimization goals; impose guardrails on reinforcement learning criteria when trust or compliance is the metric being optimized

Table 5: Type III ANOVA Results for Trust in AI

Effect	Sum Sq	Df	F value	Pr(>F)	η_p^2
Intercept	194.560	1	527.4990	$< 2.2 \times 10^{-16}$	–
context	0.918	1	2.4891	0.1155	0.02
severity	1.578	1	4.2782	0.0393^{*}	0.04
mistake_timing	7.633	3	6.8979	0.00016^{***}	0.04
context:severity	0.324	1	0.8782	0.3493	0.03
context:mistake_timing	0.636	3	0.5750	0.6318	$8.33e^{-4}$
severity:mistake_timing	4.799	3	4.3368	0.0051^{**}	0.04
context:severity:mistake_timing	0.719	3	0.6494	0.5838	$5.57e^{-3}$
Residuals	128.355	348	–	–	–

Note. (* = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$). The effect size η_p^2 (partial η^2) can be interpreted as small if $\eta_p^2 = 0.01$, medium if $\eta_p^2 = 0.06$, and large if $\eta_p^2 = 0.14$ [34].

Table 6: Means (and SD) of Trust in AI by Severity and Mistake Timing

	Mistake Timing			
Severity	**Begin**	**End**	**Never**	**Random**
High	2.98 (0.64)	2.72 (0.65)	3.40 (0.58)	3.05 (0.76)
Low	3.24 (0.59)	3.30 (0.57)	3.26 (0.55)	3.27 (0.57)

Note. Trust means (and SDs) for each combination of *Severity* and *Mistake Timing*.

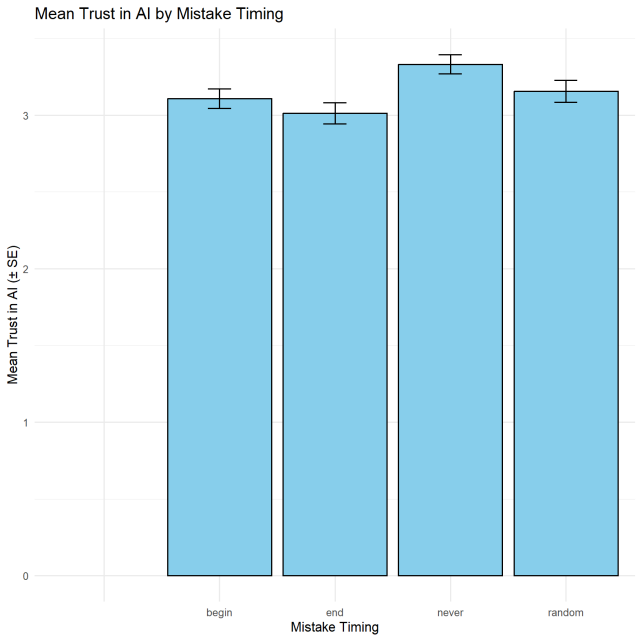


Figure 2: Mean trust across timing conditions with error bars.

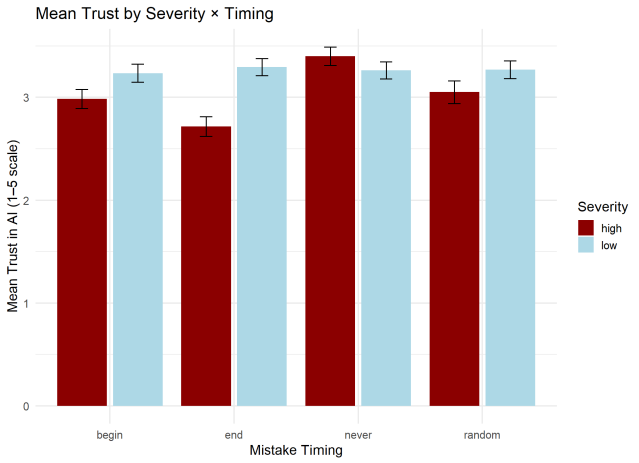


Figure 3: Mean trust across timing and severity conditions with error bars.

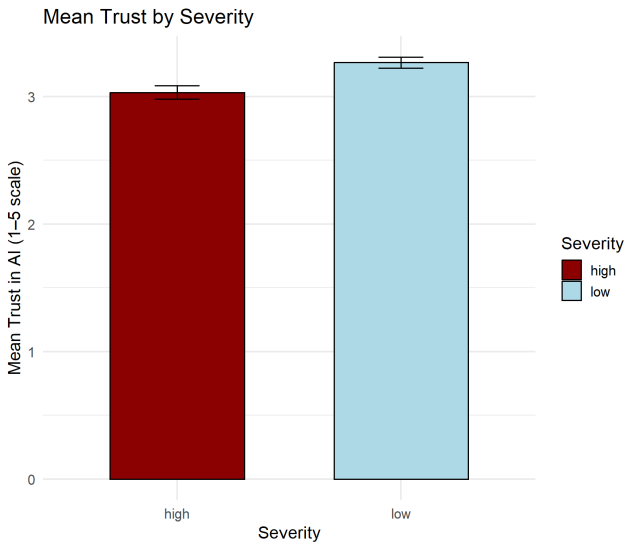


Figure 4: Mean trust across severity conditions with error bars.

A.2 THE SURVEY

Expectation Framing Prior to Task Onset

Before starting the task, participants were shown the following message:

> *'In the following experiment, an AI device is classifying images, but it may not be perfect. Therefore, we want to have you indicate whether you agree with the classification or not.'*

This framing was informed by Vereschak et al. [65], who emphasize the importance of controlling participant expectations at the start of trust experiments. Positive expectations are a prerequisite for trust formation; without them, participants may approach the system with skepticism, undermining valid trust measurement [65]. Rather than stating a fixed accuracy rate, we adopted a balanced description that acknowledges fallibility while encouraging engagement—an approach supported by prior studies Vereschak reviewed in their Trust in AI literature review within HCI. This also helped normalize error and support trust calibration across our experimental conditions.

Display the Captcha, pledge, and consent form. Then display instructions and screening questions.

- Questions about their trust in the AI device. The questionnaire and scale from the work of Gulati et al. [21] were used to measure trust in the AI device, with a Cronbach's alpha of greater than 0.84.
- Questions about their confidence and satisfaction in the AI device
- Two qualitative questions. The first question asked participants to describe the performance of the device and why they would or would not trust it. The second question asked participants to describe their initial impressions of the AI device and asked if their impression later changed and why or why not.
- Questions about cognitive load. The questionnaire and scale from the work of Hart et al. [22] were used to understand the cognitive load participants faced when interacting with the AI device.
- Question about the expertise of the participants. This questionnaire and scale from the work of Kahr et al. [27] were used to measure the level of expertise of the participants with AI.
- A question measuring self-efficacy, that asks the participant directly how confident they are in understanding the AI device.
- A question that measures how well participants think they understand AI.
- A question asking participants to rate their past experiences with AI.
- Questions about their disposition to trust humans, in general. This questionnaire and scale were taken from McKnight et al [43].
- Questions measuring personality using the big 5 personality questionnaire and scale (10 item version) [57].
- Debrief, revealing that the AI was partially simulated and had a designated 92% accuracy. This partial deception was important to test user trust in plausible high-accuracy systems.
- Five demographic questions (age, gender, race or nationality, education, income, employment, and how long they have lived in the USA.)

A 2.1 Instructions

- In the following experiment, an AI device is classifying images, but it may not be perfect. Therefore, we want to have you indicate whether you agree with the classification or not.
- First, you will be shown an image. The AI has classified the image. You will be asked if you agree with the classification. The correct classification will be shown after.
- Attention Check instructions.

A 2.2 Scenario Description

(1) Military Security Context Instructions: For the following questions, you will be reviewing images that an AI system has classified. Imagine you are a Military Operator who must report what has been flagged by an AI device for further investigation in a battlefield scenario, based on the AI system's classification of an image.
(2) Social Media Moderation Context Instructions: For the following questions, you will be reviewing images that an AI system has classified. Imagine you are a 'Social Media Moderator' regulating content for children (ages 13-17) who must make moderation decisions (such as deleting posts and images) based on the AI system's flag of inappropriate content.
(3) Annouce Main Survey

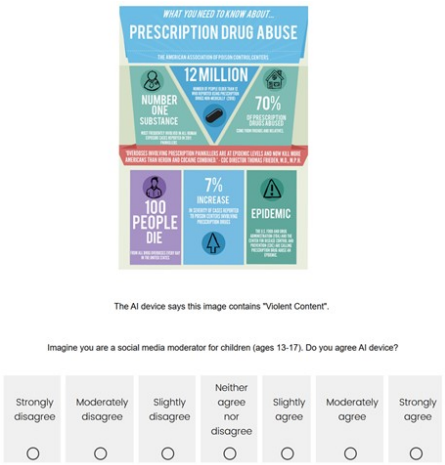


Figure 5: Incorrect classification example. Image credit: DES Daughter via Flickr, licensed under CC BY-NC-SA 2.0. This image was not part of the actual participant stimuli.

Social Media Scenario. The AI was incorrect. The previous image actually contains "Informational Awareness". This is a very minor misclassification.

Military Defense Scenario. The AI device says this is a "Cruise Ship".

Imagine you are a military operator who must make battlefield decisions based on the AI system's responses. Do you agree with the AI device?

The AI was correct. The previous image was actually a "Cruise Ship".

A 2.3 Trust Questionnaire. These questions were shown in the stated order. All responses were recorded on a Likert scale of 1-5.

- In the following questions, if you were to use the AI Device from the questions you just answered (based on how it performed), please indicate how much you agree or disagree with the following statements:
- I believe that there could be negative consequences when using (—)
- I feel I must be cautious when using (—)
- It is risky to interact with (—)
- I believe (—) will act in my best interest
- I believe that (—) will do its best to help me if I need help
- I believe that (—) is interested in understanding my needs and preferences
- I think that (—) is competent and effective in classifying images
- I think that (—) performs its role as an image classifier very well
- I believe that (—) has all the functionalities I would expect from an (—)
- If I use (—), I think I would be able to depend on it completely
- I can always rely on (—) for classifying images
- I can trust the information presented to me by (—)

A.3 Individual Differences and Covariates Measurements

Our study measured participants' prior experience with AI, trust propensity, perceived and actual AI expertise, self-reported confidence and satisfaction with the AI system, and personality traits.

A 3.1 Prior Experience with AI.

Our past experiences drive our expectations [65], and can affect the way we update our beliefs. For example, if the past experience with a system was negative, a participant is more likely to overreact to an error during an experiment, reconfirming their initial expectations [65]. Therefore, we asked participants about their prior experiences with AI, to understand any potential bias. It is recommended to assess the expertise and prior experience of users regarding both the AI-embedded systems and the task when running a study [65]. To gauge participants' baseline attitudes toward AI, we included the following question:

How would you rate your past experiences with AI?

- Positive
- Neutral
- Negative

Although this question was not drawn from a standardized questionnaire, it was designed to capture potential preexisting biases that could influence trust in AI systems.

A 3.2 AI Expertise. Participants were asked to answer questions assessing their expertise in AI and also rate their own perceived expertise with AI. In their review of trust in HCI, Vereschak et al. emphasize that vulnerability is a defining component of trust, distinguishing it from adjacent concepts like confidence. When users do not perceive a situation as risky or do not feel vulnerable to the system's actions, their reliance may reflect confidence rather than genuine trust[65]. For example, a person who assumes a system will work simply because it usually does, without considering potential risk, may exhibit confidence rather than trust, especially when the perceived stakes are low.

This distinction has practical implications for experimental design. As Vereschak et al. argue, when perceived vulnerability is low or self-confidence is high, trust assessments may be confounded by the user's belief in their own competence. In other words, people may attribute successful outcomes to their own expertise rather than to the trustworthiness of the system itself. To account for this, we controlled for participants' actual expertise and perceived AI expertise, which captures subjective confidence in one's ability to use or understand AI. This allowed us to better isolate trust as a relational construct, dependent on perceived capabilities and risks associated with the AI system, rather than as a reflection of the user's self-efficacy.

Participants were asked to rate thow much they agreed with the following questions about AI technologies on a 7-point Likert scale (1 = Strongly Disagree, 7= Strongly Agree) developed and validated by Pinski et. al. [54] This variable was included to control for baseline familiarity and potential biases in trust-related responses. Participants indicated their level of agreement with the following statements:

I have knowledge of…

(1) …the types of technology that AI is built on.
(2) …how AI technology differs from non-AI technology.
(3) …use cases for AI technology.
(4) …the roles that AI technology can play in human–AI interaction.

A 3.4 Perceived AI Expertise.

We asked participants the following questions:

(1) I am confident that I understand how to use AI to solve a problem I may have.
(7-point Likert scale: 1 = Strongly Disagree, 7 = Strongly Agree)
(2) How well do you understand Artificial Intelligence (in general, including how it works and its applications)?
(Participants selected one of the following response options:)

- No understanding
- Basic understanding of what AI is
- Good understanding of AI applications and use cases
- Strong understanding of how AI systems are built and function
- Expert knowledge in AI technologies and theories

A.4 Disposition to Trust.

We used the following scale by McKnight et al. [44] to measure participants' general disposition to trust others. The scale captures four dimensions: benevolence, integrity, competence, and trusting

stance. Participants rated their agreement with each item on a Likert scale from 1 (Strongly Disagree) to 5 (Strongly Agree).

Benevolence

(1) In general, people really do care about the well-being of others.
(2) The typical person is sincerely concerned about the problems of others.
(3) Most of the time, people care enough to try to be helpful, rather than just looking out for themselves.

Integrity

(1) In general, most folks keep their promises.
(2) I think people generally try to back up their words with their actions.
(3) Most people are honest in their dealings with others.

Competence

(1) I believe that most professional people do a very good job at their work.
(2) Most professionals are very knowledgeable in their chosen field.
(3) A large majority of professional people are competent in their area of expertise.

Trusting Stance

(1) I usually trust people until they give me a reason not to trust them.
(2) I generally give people the benefit of the doubt when I first meet them.
(3) My typical approach is to trust new acquaintances until they prove I should not trust them.

A.5 Big Five Personality.

In particular, since we are asking participants whether they agree with the AI Device or not, we should measure participant's agreeableness. We measured participants' personality traits using the 10-item version of the Big Five Inventory (BFI-10) [56]. Participants responded on a 5-point Likert scale ranging from 1 = Strongly Disagree to 5 = Strongly Agree.

Participants were asked to indicate how much they agreed with the following statements beginning with: *"I see myself as someone who…"*

(1) is reserved
(2) is generally trusting
(3) tends to be lazy
(4) is relaxed, handles stress well
(5) has few artistic interests
(6) is outgoing, sociable
(7) tends to find fault with others
(8) does a thorough job
(9) gets nervous easily
(10) has an active imagination

A.6 AI Confidence and Satisfaction.

Vereschak et al. [65] emphasize that trust is conceptually distinct from related constructs such as confidence and satisfaction, primarily due to the central role of vulnerability in trust. When vulnerability is absent—such as in low-stakes situations or when users perceive minimal risk—reliance on a system may stem from confidence rather than trust. Confidence reflects a belief in favorable outcomes without the need to evaluate alternatives or consider the possibility of failure [65]. In contrast, trust entails a willingness to accept risk despite uncertainty or potential negative consequences [59, 65].

This distinction is particularly important in human–AI interaction. If participants feel no meaningful vulnerability in using the AI system, perhaps because the task appears trivial or because they are highly familiar with similar technologies, their reliance may be based on confidence or satisfaction, not genuine trust. To disentangle these constructs, we measured and controlled for participants' confidence and satisfaction with the AI. This allowed us to assess whether trust ratings were driven by perceived trustworthiness or simply by general comfort or contentment with the interaction.

Participants answered the following items on a 7-point Likert scale (1 = Strongly Unsatisfied/Unconfident, 7 = Strongly Satisfied/Confident):

(1) How satisfied were you with the AI device?
(2) How confident are you in the AI device?

A.7 Images and Licenses

[2]All images above were *not* reproduced in the final manuscript; they were used solely as private, nonpublic stimuli for research participants. Each listing cites the presumed copyright holder or source. We rely on a *fair use* rationale under U.S. law (or similar exemptions) given the strictly educational and noncommercial setting. Any rights holders seeking further credit or removal are encouraged to contact the authors.

Table 7: Summary of Third-Party Images Used for Research Stimuli (Not Reproduced in Manuscript)[2]

Image Title / Description	Source & Date	Ownership / License	Usage & Fair Use Rationale
Weather Balloon	KIIITV news article (2022).	Likely all rights reserved (local outlet).	Shown under low-res conditions for participant analysis of a news photo. Used strictly for noncommercial, educational research; not reproduced in final paper.
Drone Image	Trevor Raab / *Popular Mechanics* (also on Yahoo! Lifestyle).	Copyright by photographer or magazine.	Participants evaluated the consumer drone depiction. Used strictly for noncommercial, educational research; not reproduced in final paper.
"Know What Drowning Looks Like"	Kenosha YMCA website (2024).	Educational content; presumed all rights reserved.	Infographic used strictly for noncommercial, educational research; not reproduced in final paper.
Suicide-Prevention Infographic	Carisk Behavioral Health & Community Care Plan (2024).	All rights reserved.	Used strictly for noncommercial, educational research; not reproduced in final paper.
"Impact/Effects of Drugs"	Department of Health, Rep. of the Philippines (2023).	Government material; treated as copyrighted.	Fair use for strictly for noncommercial, educational research; not reproduced in final paper.
"Every Child Has a Right"	The Mama Bear Effect (2021).	All rights reserved.	Displayed in a private study environment. Used strictly for noncommercial, educational research; not reproduced in final paper.
"STOP Online Sexual Harassment"	Project deSHAME/Childnet (EU Commission).	All rights reserved.	Fair use for strictly for noncommercial, educational research; not reproduced in final paper.
Human-Like Tree Trunk Photo	Unknown source (viral on Reddit/Tumblr).	Original copyright owner undetermined.	Fair use for strictly for noncommercial, educational research; not reproduced in final paper.
"Iraqi Freedom" Mural Photo	Mural by Don Gray, photo by Gerald Thurman (2003).	Artwork + photograph each copyrighted.	Fair use for strictly for noncommercial, educational research; not reproduced in final paper.
U.S. Bank Photo	*Minneapolis/St. Paul Business Journal* (2024).	Credited to U.S. Bancorp or the Journal.	Fair use for strictly for noncommercial, educational research; not reproduced in final paper.
"Cosplay Codename: KND"	DeviantArt user *knoah123* (2023).	Artist owns copyright.	Fair use for strictly for noncommercial, educational research; not reproduced in final paper.
Lake Photo	Hacche Moor Fishery website (© 2025).	Copyright by Hacche Moor Fishery.	Aerial/lake layout used for academic stimuli. Fair use for strictly for noncommercial, educational research; not reproduced in final paper.
Charity Poster	Limkokwing University (Pinterest).	Likely all rights reserved.	Fair use for strictly for noncommercial, educational research; not reproduced in final paper.
Missing Person Template	Venngage Inc.	All rights reserved.	Generic wanted-poster template for design/perception research. Resized, no commercial usage.
Minecraft Light-Up Sword	Walmart product page.	Retail promotional image, standard copyright.	Fair use for strictly for noncommercial, educational research; not reproduced in final paper.
Cruise Ship Photo	AOL / *The Independent* (2023).	News article image, likely all rights reserved.	"Icon of the Seas". Fair use for strictly for noncommercial, educational research; not reproduced in final paper.
Ducks Flying (Alamy)	Alamy stock site (license not purchased).	Watermarked stock photo, all rights reserved.	Fair use for strictly for noncommercial, educational research; not reproduced in final paper.
Epinephrine Auto-Injector Infographic	Children's Healthcare of Atlanta (2022).	All rights reserved.	Fair use for strictly for noncommercial, educational research; not reproduced in final paper.
User-Drawn Flag	Steemit user "niurkajgamboa."	Creator retains copyright.	Fair use for strictly for noncommercial, educational research; not reproduced in final paper.
Instagram Meme	@_heather_ryan_ or other user handle (faces blurred).	Typical *all rights reserved* social media.	Fair use for strictly for noncommercial, educational research; not reproduced in final paper.
Submarine Image	*The National Interest* (2023).	Publication or photographer owns rights.	Fair use for strictly for noncommercial, educational research; not reproduced in final paper.
Southwest Airlines Thumbnail	*News4SA* (Sept 13, 2024), possibly Getty.	Likely all rights reserved.	Fair use for strictly noncommercial, educational research; not reproduced in final paper.
"MG Assassins" Group Photo (X/Twitter)	User @_heather_ryan_ (Apr 26, 2019).	Poster's copyright.	Blurred faces, secure environment. Analyzed social media content under fair use.
Silhouette Group Photo	Pinterest (linked to Wattpad text).	No credited photographer, assume all rights reserved.	Fair use for strictly for noncommercial, educational research; not reproduced in final paper.
"Sushi!" Skateboarding Photo	Pinterest user "rachel <3"	Copyright likely with original photographer.	Fair use for strictly for noncommercial, educational research; not reproduced in final paper.
Fighter Jet Photo	*The Independent* (May 12, 2017), credit JOHANNES EISELE/AFP/Getty.	Getty Images license normally required.	Used watermarked version. Fair use for strictly for noncommercial, educational research; not reproduced in final paper.
Jeep Image (Polish Site)	*jedz-bezpiecznie.pl* (2021) by Maciej Kalisz.	Photographer/site retains copyright.	Fair use for strictly for noncommercial, educational research; not reproduced in final paper.
Mystery Rock Photo (Unknown Origin)	Not found in reverse-image searches; prior link is defunct	Possibly all rights reserved; exact owner/creator unverified	Fair use for strictly for noncommercial, educational research; not reproduced in final paper.
NASA Rocket (Kennedy Blog)	NASA blog (June 1, 2022) photo credited to Astra.	Possibly private contractor, not standard NASA public domain.	Fair use for strictly for noncommercial, educational research; not reproduced in final paper.
Swing Photo (Pinterest)	Pin ID 745064332094320740 by "Liane."	User retains copyright.	Fair use for strictly for noncommercial, educational research; not reproduced in final paper.
Statue (Foursquare)	User "Carl J." on Foursquare (Nov 6, 2015).	Photo presumably copyrighted to user.	Fair use for strictly for noncommercial, educational research; not reproduced in final paper.

All images above were *not* reproduced in the final manuscript; they were used solely as private, nonpublic stimuli for research participants. Each listing cites the presumed copyright holder or source. We rely on a *fair use* rationale under U.S. law (or similar exemptions) given the strictly educational and noncommercial setting. Any rights holders seeking further credit or removal are encouraged to contact the authors.