INFO 300A: Research Methods AUT 2025

Research Proposal Final Submission - Group BA 3

Human Judgment vs. AI Alerts: Exploring Trust in Automated Scam Detection Systems

Joseph Chamdani, Winson Teh, Kenneth Wu, Kunxing Zeng

University of Washington

jchamd@uw.edu, wteh@uw.edu, knw2004@uw.edu, kunxing@uw.edu

Abstract: Artificial intelligence systems increasingly provide scam alerts inside email and messaging platforms, yet users often struggle to decide whether to trust these warnings. This study investigates how people balance their own judgment against AI scam alerts in fast-paced decision environments. We will conduct a controlled online experiment where participants review simulated messages that vary by scam likelihood, alert type, and time pressure, followed by short qualitative responses about their reasoning. The study measures user trust, compliance with alerts, and how these outcomes differ across alert design and age groups. We expect that explanatory alerts will produce higher trust and better decision accuracy than generic alerts. We also anticipate that time pressure will reduce careful evaluation and lead to more errors, especially among older adults. The goal of this research is to understand when users trust or dismiss AI scam alerts, and which factors most strongly influence their decision-making in real-time scam situations.

1 INTRODUCTION

As artificial intelligence has become integrated into everyday technology, automated scam detection systems have increasingly appeared in email clients, messaging apps, and smartphones. These AI-powered systems work to protect users by flagging suspicious emails, detecting fraudulent phone calls, and warning about potential phishing attempts. At the same time, scammers themselves have begun using AI tools to create more sophisticated attacks, with platforms like WormGPT and FraudGPT enabling them to generate convincing scam messages at scale (Ahmadi, 2023).

Within this evolving landscape of AI-powered attacks and defenses, research has generally focused on the technical capabilities of detection systems, investigating topics such as machine learning algorithms for fraud identification (Mathew, 2025), the application of natural language processing to analyze scam patterns (Papasavva et al., 2025), and how AI agents can simulate realistic scam scenarios (Badhe, 2025). These systems, designed to protect users from financial and psychological harm, offer unique challenges because their effectiveness depends entirely on whether people actually trust and follow their warnings.

Investigating how users navigate the tension between automated alerts and their own judgment is crucial for understanding whether these protective technologies actually work in practice. This paper uses existing scholarship to ask the following questions: How do individuals balance their own judgement with AI-generated scam alerts? How do different alert designs influence trust and accuracy? And how do age and technology experience shape the ways users interpret and act on these alerts?

2 SYNTHESIS OF EXISTING SCHOLARSHIP

2.1 The Landscape of AI-Powered Fraud

The rapid advancement of artificial intelligence has fundamentally transformed the digital threat landscape, with AI-generated scams emerging as one of the most pressing cybersecurity challenges of our time. They represent a significant threat to individuals, businesses, and society as a whole, as scammers increasingly make use of sophisticated machine learning models to automate and scale their operations (Leong et al., 2024).

Leong and colleagues describe AI as a "double-edged sword," where the same technologies designed to enhance productivity can be weaponized to perpetrate fraud with alarming efficiency. Scammers now employ generative adversarial networks (GANs) to create fake news articles and fraudulent advertisements, deep learning text generation models to craft convincing phishing emails, and voice synthesis technologies such as WaveNet and Tacotron to generate human-like speech for voice phishing attacks (Leong et al., 2024). The automation of these techniques allow a single scammer to simultaneously target thousands of victims with individually tailored deceptions, fundamentally changing the economics and effectiveness of fraudulent schemes.

Recent research shows that AI-powered scams have evolved from simple automated messages to systems that can carry out long, convincing conversations. Traditional online scams relied on mass-email campaigns that sent identical messages to thousands of victims (Alkhalil et al., 2021). Badhe (2025) introduced ScamAgent, an autonomous multi-turn agent that maintains dialogue memory, adapts to victim responses, and uses persuasive strategies across multiple turns (Badhe, 2025). When paired with modern text-to-speech systems, these AI-generated scripts create a fully automated scam pipeline requiring minimal human oversight (Badhe, 2025). This technology also supports threat vectors such as deepfake impersonation, AI chatbots used for social engineering, and AI-driven profiling that crafts highly targeted attacks (Leong et al., 2024). As a result, people face risks such as identity theft, privacy breaches, and manipulation through AI-powered scams that appear highly credible at scale (Leong et al., 2024).

Businesses can also face substantial financial losses and reputational damage through social engineering attacks such as business email compromise scams, where attackers impersonate executives or vendors to authorize fraudulent transactions, while fake content generated by AI algorithms erodes consumer trust with fraudulent advertisements and reviews. At the societal level, AI-driven scams undermine confidence in digital interactions and online platforms, with social media campaigns spreading misinformation and manipulating public opinion (Leong et al., 2024).

The democratization of AI technologies enables malicious actors with limited resources to orchestrate large-scale fraud operations. This evolution creates an imperative for sophisticated detection mechanisms, while raising fundamental questions about whether detection technology can keep pace with rapidly evolving threats and, more critically, whether users will trust and act upon their warnings.

2.2 Understanding Trust in Automated Systems

The study of trust in automation has evolved significantly across multiple disciplines including psychology, human factors engineering, and human-computer interaction. Trust in automated systems represents a person’s willingness to rely on system recommendations based on their beliefs about the system’s capability, reliability, and appropriateness for specific tasks. This psychological commitment extends beyond simple compliance, requiring users to genuinely believe the system will perform as promised and that following its recommendations serves their best interests.

When people interact with automated systems, two opposite trust patterns often appear. The first is automation bias, where users rely too heavily on system outputs without checking for supporting evidence. The second is algorithm aversion, where users ignore or undervalue correct system recommendations because they do not trust the technology. Trust is strongly influenced by error sensitivity. A single mistake from a system perceived as reliable can quickly reduce confidence, while users are more forgiving toward systems they already believe are imperfect. People also do not accept or reject AI advice uniformly. Instead, they choose when to rely on or ignore automated suggestions based on the situation and how confident they feel in their own judgment.

The best kind of human AI interaction happens when people use AI for tasks it does well but still rely on their own judgment when human context or common sense is needed. Several things affect how people build trust in these systems, such as transparency, accuracy, and past experience. For instance, a warning that says "This email looks suspicious because it asks for your password and comes from a mismatched domain" feels more trustworthy than a vague "This is suspicious." Accuracy also plays a role, though people's perception of reliability doesn't always match reality, and their past experience and comfort with technology influence how much they trust it.

Demographic factors, particularly age, substantially influence trust in AI detection systems. Research examining AI tools for identifying fraudulent online reviews found that younger adults (ages 20-26) demonstrated significantly higher trust in AI judgments compared to older adults (ages 50-78), who exhibited greater skepticism toward automated assessments (Xiang et al., 2022). This demographic variation suggests that uniform warning design may prove ineffective across diverse user populations.

2.3 The Unique Challenge of Scam Detection

AI systems face special challenges in detecting scams because scammers often change their methods to trick detection tools. While many AI applications use steady or predictable data, scam detection must deal with human deception that happens in real time. Hossain et al. (2025) explain that effective scam detection needs flexible, privacy-focused models that can analyze live conversations instead of only stored information. For their study, they developed a real-time monitoring framework using federated learning to detect emotional cues and manipulative language while keeping user data private. Their results showed that their model could identify scam interactions with strong accuracy and continuously improve its performance through feedback from distributed devices. This outcome reflects a broader shift in cybersecurity, where detection tools are no longer static filters that only flag known threats after they occur, but active systems that learn, adapt, and respond to new scam behaviors as they emerge.

In scam situations, people are often forced to make decisions in seconds under fear, pressure, or confusion created by the scammer. Because of this, users may ignore AI warnings and trust their instincts instead, or they may hesitate because they do not fully trust the system. Hossain et al. (2025) point out that good detection systems must balance how strongly they intervene with how much freedom they allow the users to have, since too many interruptions can reduce confidence. The authors argue that AI warning systems should focus not just on being technically accurate but also on being easy to understand and trustworthy, so that users will take their warnings seriously.

Papasavva et al. (2025) found that users’ responses to scam alerts depend strongly on how warnings are framed and contextualized. Their study showed that alerts combining visual cues, such as red banners or warning icons, with short explanations about why a message was flagged led to higher trust and faster recognition compared to warnings that only displayed generic caution labels. Together, these studies show that successful scam detection depends as much on user-friendly design as on advanced algorithms. Future research should look more closely at how people notice, understand, and react to AI scam alerts when they are under stress and must decide quickly.

2.4 Summary

Artificial intelligence is used in many fraud detection systems, but people still do not always know how reliable these tools are or when to trust them. While AI can flag suspicious messages, Mathew and Fofang (2025) showed that better accuracy does not automatically lead users to follow the alerts. Most research focuses on improving the technology itself instead of understanding how people react to these warnings.

Because of this, there is still a gap in what we know. We do not fully understand how people decide to trust or ignore AI scam alerts, especially when they feel rushed or uncertain. There is also limited research on how alert design and individual factors, like age or technology experience, affect these decisions.

Therefore, the proposed study seeks to fill this gap in the literature by addressing the following research questions:

1. How do individuals balance their own judgement with AI-generated scam alerts during digital interactions?

2. How do different alert designs, such as generic warnings versus explanatory warnings, affect user trust and decision accuracy?

3. How do age and technology experience shape user responses to AI scam alerts?

3 Research Method and Methodology

This research proposal examines how individuals balance their own judgment against AI-generated scam alerts during fast-paced digital interactions. To investigate this behavior, the study will adopt a mixed-methods experimental approach, integrating both quantitative and qualitative data. This methodology is appropriate because trust, decision-making, and user interaction with AI involve not only measurable behavioral outcomes (such as accuracy, compliance, or response time) but also subjective experiences that require participants to explain *why* they trusted or ignored an alert. Compared to a purely qualitative design, a mixed approach allows the study to capture both what decisions users make and why they make them. A fully quantitative approach was rejected because it cannot adequately capture the cognitive and emotional reasoning behind trust in AI warnings.

3.1 Methodological Standpoint

This proposal follows a post-positivist epistemological standpoint, which acknowledges that human behavior can be studied through observable patterns, but that people's judgments, decisions, and trust in technology are also shaped by subjective interpretation. Trust in AI alerts is not purely objective, so combining measurable behavioral outcomes with self-reported reasoning aligns with a post-positivist perspective. The theoretical perspective guiding the study draws from human–computer interaction and trust in automation theory, which emphasize the interaction between system design, user expectations, and perceived reliability.

3.2 Sample and Population

The study will recruit approximately 60–80 participants, focusing on two demographic groups:

- Younger adults (ages 18–30)
- Older adults (age 50+)

Previous research shows age strongly impacts trust in automated systems, making demographic comparison important. This study focuses specifically on younger adults and older adults because these groups show the greatest differences in trust and reliance on AI-based alerts. Adults between ages 30 and 50 are not included because prior research has not identified clear trust differences for this middle group, and this study aims to focus on age ranges where patterns are most distinct. Participants will be recruited through university mailing lists, community centers, and online postings. A purposive sampling method will be used to ensure representation across the two age groups. All participants must own a smartphone and regularly use email or messaging apps.

3.3 Data Collection

Data will be collected using a controlled online experiment administered through a custom browser-based interface. Participants will be shown a series of short scam-like messages (e.g., phishing emails, suspicious texts, fake financial warnings) that vary by:

- Scam type (phishing, urgency, impersonation),
- Alert style (generic vs explanatory AI warning),
- Time pressure (low vs high)

During each trial, participants will :

1. Read the message
2. View an AI-generated scam alert
3. Decide whether to trust the alert or rely on their own judgement
4. Select their reasoning from a list and optionally provide a short explanation

3.4 Data Analysis

For the analysis, both the numerical results and the short explanations from each participant will be examined. The quantitative section will compare how often participants correctly identify scams, how often they follow the AI alert, and how long they take to respond in each scenario. Average results will be compared across groups to determine which factors increase accuracy or influence trust in the alert.

For the qualitative section, the short written responses will be reviewed to identify common themes in how participants described their decisions. This analysis will clarify the reasons some individuals trusted the alert while others chose to ignore it.

3.5 Threats to Validity

There are several threats to validity that may affect this study. The first threat is hypothesis guessing, which is a type of construct validity threat. Participants may realize that the study is focused on trust in AI scam alerts and may change their behavior to match what they think the researcher expects. This would reduce the accuracy of the findings. To reduce this threat, the study description will not reveal the focus on trust or decision making.

A second threat is selection bias, which affects internal validity. Participants are recruited from university and community groups, and this may result in a sample that has higher digital literacy than the general population. If participants are more comfortable with technology, this could influence how they interpret scam messages. To address this issue, the study will collect basic information about participants' experience with technology and will compare responses across different experience levels.

A third threat concerns ecological validity, which is a type of external validity. Participants respond to simulated scam messages rather than real ones. Real scam situations often involve emotional pressure or urgency that is difficult to recreate. This limits how well the results can be generalized to real settings. To reduce this issue, the study will use realistic message designs and will vary the urgency and style of each scenario.

4. Discussion

The expected findings of this study will help explain how people balance their own judgment with AI-generated scam alerts and which types of warnings lead to the most accurate decisions. Prior research suggests that alerts that briefly explain why a message appears suspicious can increase trust and accuracy, while generic alerts may be ignored. Younger adults may rely more on automated guidance, while older adults may remain cautious, especially when making decisions under time pressure. These differences are important because they can show which groups benefit the most from improved alert design and how warning systems should adjust to different levels of digital experience.

Beyond expected results, the broader significance of the study lies in improving how scam detection systems support users in real-world situations. The findings can help technology companies and cybersecurity designers build clearer and more trustworthy alerts. The study also includes several challenges and limitations. Simulated messages cannot fully capture the emotional pressure of real scam encounters, and participants from university settings may demonstrate higher digital literacy than the general population. Ethical considerations include avoiding unnecessary distress and ensuring that participants understand the scenarios during the debriefing. Data collection may also be affected by rushed responses or inconsistent device conditions. Future issues could involve the rapid evolution of AI-generated scams, which may require updated research as scam tactics change. Results from this study could be shared through academic publications, HCI conferences, and organizations that focus on consumer digital safety.

5. Conclusion

This research proposal outlines a study examining how people respond to AI-generated scam alerts and how they balance these alerts with their own judgement in fast-paced digital situations. Although technical methods for scam detection have improved, much less is known about how users interpret and act on these warnings. This study addresses that gap by observing how alert design, time pressure, and demographic factors influence trust and decision making.

Using simulated messaging scenarios and both quantitative and qualitative data, the study will show why some users rely on AI alerts while others ignore them. These findings can help designers create clearer and more effective scam warnings and support safer digital environments. They can also guide future improvements in email platforms, messaging apps, and other digital tools.

Overall, this proposal emphasizes the importance of user trust in automated detection systems and offers a useful approach for understanding how people make decisions during real-time scam situations.

REFERENCES

Ahmadi, S. (2023). Open AI and its impact on fraud detection in financial industry. *Journal of Knowledge Learning and Science Technology, 2(3), pp.263-281.* https://hal.science/hal-04456232/

Xiang, H., Zhou, J., & Xie, B. (2022). AI tools for debunking online spam reviews? Trust of younger and older adults in AI detection criteria. *Behaviour & Information Technology, 42*(5), 478–497. https://doi.org/10.1080/0144929X.2021.2024252

Leong, W. Y., Leong, Y. Z., & Leong, W. S. (2024). The intersection of scammers and artificial intelligence. In *2024 International Conference on Consumer Electronic - Taiwan (ICCE-Taiwan)*. https://doi.org/10.1109/ICCE-Taiwan62264.2024.10674334

Alkhalil, Z., Hewage, C., Nawaf, L., & Khan, I. (2021, March 8). Phishing attacks: A recent comprehensive study and a new anatomy. *Frontiers.* https://doi.org//10.3389/fcomp.2021.563060

Badhe, S. (2025, August 8). Scamagents: How AI agents can simulate human-level scam calls. *arXiv.org*. https://arxiv.org/abs/2508.06457

Papasavva, A., Lundrigan, S., Lowther, E. et al. Applications of AI-Based Models for Online Fraud Detection and Analysis. *Crime Sci* **14**, 7 (2025). https://doi.org/10.1186/s40163-025-00248-8

Hossain, I., Puppala, S., Talukder, S., & Alam, M. J. (2025, September 12). *Ai-in-the-Loop: Privacy Preserving Real-Time Scam Detection and Conversational Scambaiting by Leveraging LLMs and Federated Learning. arXiv.org*. https://arxiv.org/abs/2509.05362

Mathew, A., & Fofang, T. (2025). AI-Driven Fraud Detection: Leveraging Machine Learning for Scam Identification. *International Journal of Innovative Research in Science, Engineering and Technology*, *14*(04). https://doi.org/10.15680/ijirset.2025.1404002

6 CHATGPT STATMENT

We did not use ChatGPT for this assignment. However, we did use ChatGPT prior to this assignment to narrow our focus and generate ideas for our study during the beginning stages of our research. We also took the draft to a writing tutor for feedback and proofreading to check for clarity, consistency, and proper APA formatting.