Position: Human Factors Reshape Adversarial Analysis in Human-AI Decision-Making Systems

Shutong Fan
shutonf@clemson.edu
Clemson University

Lan Zhang
lan7@clemson.edu
Clemson University

Xiaoyong Yuan
xiaoyon@clemson.edu
Clemson University

Abstract

As Artificial Intelligence (AI) increasingly supports human decision-making, its vulnerability to adversarial attacks grows. However, existing adversarial analysis predominantly focuses on fully autonomous AI systems, where decisions are executed without human intervention. This narrow focus overlooks the complexities of human-AI collaboration, where humans interpret, adjust, and act upon AI-generated decisions. Trust, expectations, and cognitive behaviors influence how humans interact with AI, creating dynamic feedback loops that adversaries can exploit.

This position paper argues that **human factors fundamentally reshape adversarial analysis in human-AI decision-making systems.** We first revisit Human-AI collaborative systems through the lens of adversarial robustness and discuss how human factors create new attack surfaces beyond conventional model-centered adversarial examples. We then introduce a novel robustness analysis framework that (1) examines how human factors affect collaborative decision-making performance, (2) interprets existing adversarial attack strategies in the context of human-AI interaction. Our goal is to reconceptualize adversarial robustness as a non-negligible property of the coupled human-AI system, where vulnerabilities are shaped not only by model behavior but also by evolving human reliance, adaptation, and decision-making dynamics. An extended version of this paper is available at: https://arxiv.org/pdf/2509.21436.

Workshop on Designing Human–AI Teaming for Security (HATS), co-located with the Symposium on Usable Privacy and Security (SOUPS) 2026.
August 23, 2026, Hannover, Germany.

1 Motivations

With rapid advancements in artificial intelligence (AI), an increasing number of tasks across diverse domains are performed with unprecedented efficiency and accuracy, benefiting human decision-making. For example, AI assists drivers in navigating complex road conditions [12], and helps professionals refine documents with tools such as ChatGPT [25]. Despite these advancements, AI remains highly vulnerable to adversarial attacks - subtle input perturbations designed to mislead models while remaining imperceptible to humans [35]. Such attacks pose significant risks to AI-enabled decision-making, particularly in safety-critical applications.

While extensive research has examined adversarial attacks on fully automated AI systems, it often overlooks a crucial factor: human involvement in AI-assisted decision-making systems. In real-world applications, humans provide oversight, make critical judgments, and address broader societal considerations in AI deployment [24, 28]. We argue that **human involvement fundamentally reshapes adversarial attack strategies and redefines robustness analysis in human-AI decision-making systems**. Prior work has investigated how to mislead autonomous driving systems, creating serious safety hazards [16]; however, human intervention can alter outcomes: humans may recognize anomalies in driving decisions and take corrective action, mitigating potential accidents [27]. Unfortunately, little attention has been paid to how adversarial attacks can strategically exploit human cognitive and behavioral tendencies to influence or distort decision-making processes. Addressing these gaps is critical for developing AI systems that are not only resilient to adversarial threats but also aligned with human intuition and reasoning.

Within the limited research that incorporates humans into adversarial analysis, most studies implicitly treat humans and AI models as independent entities, evaluating their responses to adversarial inputs in isolation [26]. For instance, adversarial examples have been crafted to deceive both AI systems and time-constrained humans [31]. These approaches primarily

evaluate whether adversarial inputs can fool both human and AI systems in parallel, but critically, they assume that human decisions are made independently of prior AI predictions and that human feedback does not influence subsequent AI behavior. However, they overlook a critical aspect of real-world collaborative deployments, where humans rely on, interpret, and adapt to AI outputs over time. Ignoring this interdependence limits the understanding of how adversarial attacks propagate through human-AI teams and how trust dynamics evolve under adversarial pressure.

To bridge this gap, **we advocate a comprehensive exploration of human factors in adversarial analysis, to enhance the robustness of human-AI decision-making systems.** Rather than acting as isolated variables, human factors, such as self-confidence, dynamically shape reliance on AI, which directly influence when and how users choose to rely on AI outputs and, in turn, determine the effectiveness of adversarial attacks. Understanding these interactions is essential for identifying system vulnerabilities and designing defenses.

2 Revisiting Human-AI Collaborative Systems Through Lens of Adversarial Robustness

Adversarial examples have been recognized as a fundamental vulnerability in modern AI systems [13]. Most prior work focuses on perturbing inputs to fool a model in fully autonomous settings [7]. These attacks typically assume a static decision-maker and evaluate success solely by changes in model predictions. However, in many real-world applications (e.g., autonomous driving, medical diagnosis), humans remain in the loop as ultimate decision-makers. In such settings, adversarial impact is not limited to model misprediction; it also depends on how humans perceive, evaluate, and respond to AI outputs.

Analyzing adversarial robustness in human-AI collaborative systems, therefore, requires looking beyond model performance. Cognitive and situational factors in humans can shape when users rely on AI, when they override it, and how prior AI behavior affects future trust. The key factors include self-confidence, task risk, task complexity, and time sensitivity, which turn human reliance into a dynamic component of the attack surface.

Building on these insights, we present four observations about how adversarial robustness changes in human-AI decision-making systems:

Obs. 1: Model performance is the primary anchor of human trust. Humans gain more trust in an AI system as its predictions become more accurate [34]. Even small, adversary-induced drops in observed performance (e.g., accuracy) will erode that trust: after observing just a few conspicuous errors, users often hesitate to use the AI system on subsequent tasks [20].

Since trust is performance-driven, the adversary attack does more than create mispredictions; it also signals unreliability, prompting users to disengage from future AI assistance. The stronger or more frequent the perturbations, the faster confidence collapses; once users stop relying on the model, the attack can no longer influence decisions, resulting in a self-limiting trust–performance feedback loop.

Insight 1: Adversarial attacks trigger a self-limiting trust–performance feedback loop.

Obs. 2: Human reliance on AI is multifaceted. The growing human-AI interaction research show that human reliance on AI is multiplex [33]. Model performance alone cannot explain when people accept or reject AI. Prior work identified additional cognitive and situational factors:

Cognitive factors: (1) *Self-confidence* refers to the user's belief in their own judgment. This trait often varies by domain; individuals tend to exhibit higher self-confidence in areas where they possess expertise and lower confidence in unfamiliar domains [22]. Researchers state that individuals with high self-confidence are less inclined to defer to AI, even when the system is highly certain [8].

Situational factors: (1) *Task risk* captures the safety-critical aspect of a task. For instance, in autonomous driving scenarios, driving through heavy rain at night carries a significantly higher risk than driving on a clear, quiet road. Prior works find that the same decision-makers become more willing to use AI when task risk is high [11, 18]. (2) *Task complexity* reflects the inherent difficulty of the task. For instance, solving calculus problems typically involves higher complexity than performing basic arithmetic operations. Existing findings on the relationship between task complexity and user trust remain mixed: some studies report higher trust for simple tasks [5], others for complex tasks [9], and some observe a U-shaped pattern with the lowest trust at moderate complexity [38]. Recent work suggests that different dimensions of complexity (components, coordination, dynamics) may underlie these discrepancies [29]. (3) *Time sensitivity* is the urgency of a task or the speed required for decision-making. For example, autonomous driving demands rapid responses, whereas using a large language model for literary translation allows more deliberation. Longer decision windows allow users to verify AI suggestions, while they tend to rely on AI outputs under tight deadlines with less scrutiny [6].

These factors modulate human reliance, thereby reshaping the landscape for adversarial analysis. Therefore, adversarial attacks should consider more than just optimizing perturbations against AI models. The cognitive and situational factors listed above also play critical roles in determining attack effectiveness and strategies. We consider three tasks: (1) **Low-risk tasks** (e.g., e-mail triage, movie picks). Users often accept model suggestions without scrutiny, lowering the chance that subtle manipulations or minor errors will be noticed. In such cases, adversarial perturbations can be more significant, allowing attackers to increase attack frequency without detection. (2) **High-risk, slow-paced tasks**

(e.g., financial approval, non-urgent clinical reviews). The users inspect model output more closely when mistakes are costly [15]. An attacker must craft subtler perturbations. Once users detect mistakes, they may lose trust and abandon the AI model entirely, limiting the impact of the adversarial attacks. (3) **Time-critical tasks** (e.g., autonomous driving, real-time trading). Under time constraints, even in high-risk scenarios, users may bypass additional verification and rely entirely on the AI's output [36], creating a window of vulnerability that attackers can exploit.

These scenarios show that human factors other than model performance can dampen or magnify adversarial effects, create new attack paths. Adversarial analysis must, therefore, consider both cognitive and situational factors.

Insight 2: Cognitive and situational factors shift where, when, and how an attacker can succeed.

Obs. 3: Humans take actions in decision-making; the reliance shapes human actions. Humans actively participate in the decision-making pipeline; they are not passive recipients of AI output [37]. Once trust is established, it drives concrete human behavior. In the decision-making loop, humans continually determine whether to watch the AI, when to step in, and how to correct it [15]. The amount of trust humans have in the AI system triggers different actions. For instance, clinicians may choose to pay more attention to the AI suggestions and tend to reject the suggestions when they detect frequent inconsistencies with patient context or prior knowledge [5, 30].

These meta-decisions introduce an additional layer of defense, creating robustness that adversarial attacks must circumvent. An attack succeeds only if the incorrect predictions still result in the final decision. Any adversarial analysis that ignores these human safeguard actions will overstate how much damage an attacker can actually cause.

Insight 3: Adversarial attacks that fool the AI model are only halfway successful.

Obs. 4: Human-AI decision-making systems are interactive and unfold over time. Unlike AI-only systems that often treat each decision as isolated, human-AI collaboration is sequential: past interactions shape future reliance and decisions [20, 32]. Users update their reliance after each success or failure; then, reliance determines their actions, e.g., how closely they will monitor, accept, or override the model in the following cases. This feedback loop turns even "single-step" tasks (e.g., bail decisions, medical diagnoses) into a sequential process: a judge who consults the AI across several hearings, or a clinician who revisits an AI assistant across patients, accumulates experience that shifts future reliance on AI. The result is a moving target for both accuracy and vulnerability, different from an isolated AI system.

The temporal structure of human-AI decision-making introduces new attack surfaces: adversaries can not only target individual predictions but also manipulate the dynamics of human trust. If the model is accurate for the first few tasks, users relax their checks. A significant error can then pass unnoticed, transforming a minor perturbation into a high-impact failure. Consequently, the objectives for both attackers and defenders should shift from success in isolated tasks to cumulative effectiveness over entire interaction sequences.

Insight 4: Adversarial analysis must consider the full sequence of interactions, rather than a single isolated decision point.

3 Robustness Analysis Framework for Human-AI Decision-Making

3.1 Overview

To characterize the evolving human role in AI-interactive decision-making, we introduce a robustness analysis framework that models human behavior under adversarial conditions. Inspired by the four observations in Section 2, our framework consists of two assessments: **AI reliance assessment** and **AI performance assessment**, as shown in Figure 1. The **AI reliance assessment** captures how users decide whether to trust AI predictions based on both prior AI performance and model-irrelevant factors, which are task-specific cognitive and situational factors (e.g., self-confidence and task complexity). The **AI performance assessment** evaluates whether an AI output is reliable through human feedback, then affects future trust and governs whether to execute, override, or fall back. Together, the two assessments capture how users interact with AI predictions, update their trust, and respond to adversarial threats within this temporally sequential setting.

3.2 Assessment 1: AI Reliance

AI reliance assessment determines whether humans choose to trust an AI model's prediction. As discussed in Section 2, this decision is shaped by two primary components: (1) AI performance feedback, reflecting prior experiences with the model, and (2) AI model-irrelevant factors, including the cognitive and contextual factors (e.g., task risk, self-confidence).

- *AI performance feedback.* AI performance captures historical performance trends, which dynamically affect future reliance in human-AI collaboration. Analogous to human-human collaboration, unsatisfactory experience lowers the likelihood of partnering [1, 4]. Human-AI collaboration follows the same pattern that dissatisfying outcomes reduce willingness to rely on the AI model, while consistent successes rebuild trust.
- *AI model-irrelevant factors.* These include cognitive and situational variables such as self-confidence and task complexity, which influence trust independently of the AI's actual performance (discussed in Obs. 2). In addition to considering the AI's historical performance, users evaluate the model-irrelevant factors to ultimately decide whether to adopt the

AI's predictions. For example, users may reject AI predictions in high-risk tasks such as surgery and air traffic control, even when the system has perfect historical performance [3].

When reliance conditions are met, humans defer to the model without making their own prediction. Otherwise, they generate an independent decision. Existing work has extensively examined trust in AI [2, 17], yet most studies overlook the role of AI performance in shaping trust. Our framework addresses this gap by incorporating historical AI performance as a dynamic factor, allowing trust to evolve over time.

3.3 Assessment 2: AI Performance

AI performance assessment evaluates the acceptability and reliability of AI predictions in two scenarios: when humans (1) do not trust or (2) trust AI.

- *When humans do not trust AI (low reliance in Assessment 1)*: When humans do not defer to the AI, they generate their own prediction and compare it with the AI output, referred to as human-model prediction assessment. Evaluation methods vary from binary correctness checks (match/mismatch) to metric-based assessments (e.g., BLEU [23]). For example, a driver with low reliance on an autonomous driving system may independently verify a stop sign and intervene if the vehicle fails to stop.
- *When humans trust AI (high reliance in Assessment 1)*: In most cases, humans adopt the AI's predictions directly, resulting in the AI decision execution. However, they may still assess predictions indirectly through downstream behavior, known as model-only prediction assessment. For example, a driver might not visually inspect a manipulated stop sign but would notice unexpected acceleration and intervene.

Assessment 2 serves two primary functions. One is *decision execution*. If AI's prediction is deemed reliable, it is executed. Otherwise, human intervention overrides the AI decision when a human prediction is available. If no available human prediction (i.e., AI model is trusted), a failsafe or fallback mechanism is triggered. Another one is *feedback loop*. Performance evaluation results inform future AI reliance assessments, ensuring trust is dynamically updated based on AI reliability. Consistent with prior work [17], trust in AI increases with positive experiences and decreases with observed failures, leading to more intensive human assessment.

3.4 Human-Aware Adversarial Attack

Building on the dual assessments of AI reliance and performance introduced in our framework, we now examine how adversarial strategies must adapt in the context of human-AI collaboration. Traditional adversarial attacks target AI-only systems, where models execute predictions without human oversight. However, in human-AI interactive systems, attackers must also account for human monitoring, intervention,

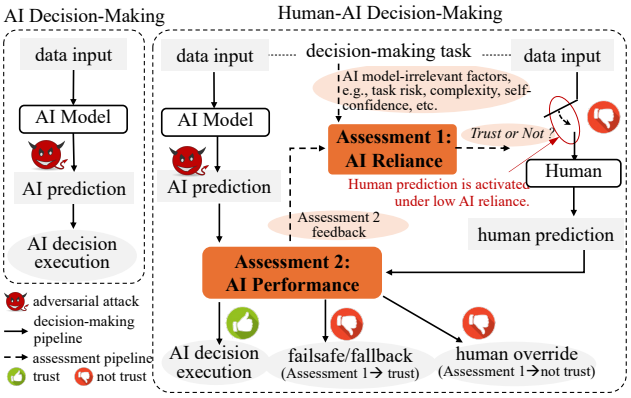

Figure 1: Comparison between conventional AI-only decision-making (left) and our proposed robustness analysis framework for human-AI decision-making (right).

and evolving trust. Our framework introduces two human-aware adversarial considerations: 1) *Human Awareness of Anomalies*: Unlike fully automated pipelines, humans may detect adversarial perturbations even without direct input observation. For example, in autonomous driving, a driver may not see a manipulated stop sign but may notice unexpected acceleration and override the AI. 2) *Adaptive Trust Exploitation*: Adversaries must strategically balance the impact of their attacks with the need to avoid rapid trust degradation to sustain long-term influence. Instead of consistently triggering obvious errors that erode human reliance on AI, attackers may selectively introduce perturbations, adapting to the user's evolving trust to maintain influence over decision-making.

4 Open Challenges and Future Directions

This position paper highlights how human dynamics can reshape adversarial analysis in human-AI decision-making systems. However, incorporating human factors into robustness analysis raises several open challenges to be studied.

- *Realistic human behavior modeling.* Our pipeline abstracts reliance using factors such as self-confidence and task risk, but real-world reliance is more heterogeneous and context-dependent [19]. Users differ in expertise, prior experience, and willingness to override AI outputs [21]. Future work should develop adaptive, empirically grounded models of reliance under adversarial conditions.
- *Robustness-usability trade-offs.* Defenses such as warnings, uncertainty displays, and fallback mechanisms may reduce adversarial risk, but can also slow decision-making or reduce user satisfaction. Overly cautious systems may further reduce trust in AI outputs [10]. Future defenses should be context-aware, activating when manipulation is likely while preserving smooth interaction during normal use.
- *Group-level human-AI collaboration.* Many real-world decisions involve teams rather than single users. In settings such as clinical teams or organizational decision-making,

reliance may be socially reinforced, contested, or redistributed across stakeholders [14]. Extending adversarial analysis from individual reliance to group-level dynamics is important for understanding collaborative human-AI security.

Acknowledgments

This material is based upon work supported in part by the U.S. National Science Foundation under Grant No. 2616640.

References

[1] Gaurav Bansal and Fatemeh Mariam Zahedi. Trust violation and repair: The information privacy perspective. *Decision Support Systems*, 71:62–77, March 2015.

[2] Michaela Benk, Sophie Kerstan, Florian von Wangenheim, and Andrea Ferrario. Twenty-four years of empirical research on trust in AI: A bibliometric review of trends, overlooked issues, and future directions. *AI and Society*, 40:2083–2106, 2025.

[3] Yochanan E Bigman and Kurt Gray. People are averse to machines making moral decisions. *Cognition*, 181:21–34, 2018.

[4] Iris Bohnet and Richard Zeckhauser. Trust, risk and betrayal. *Journal of Economic Behavior & Organization*, 55(4):467–484, 2004.

[5] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), April 2021.

[6] Shiye Cao, Catalina Gomez, and Chien-Ming Huang. How time pressure in different phases of decision-making influences human-AI collaboration. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2), October 2023.

[7] Amirhosein Chahe, Chenan Wang, Abhishek Jeyapratap, Kaidi Xu, and Lifeng Zhou. Dynamic adversarial attacks on autonomous driving systems. *arXiv preprint arXiv:2312.06701*, 2023.

[8] Leah Chong, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior*, 127:107018, 2022.

[9] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126, 2015.

[10] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. The role of trust in automation reliance. *International journal of human-computer studies*, 58(6):697–718, 2003.

[11] Hannah Fahnenstich, Tobias Rieger, and Eileen Roesler. Trusting under risk – comparing human to AI decision support agents. *Computers in Human Behavior*, 153:108107, 2024.

[12] Divya Garikapati and Sneha Sudhir Shetiya. Autonomous vehicles: Evolution of artificial intelligence and the current industry landscape. *Big Data and Cognitive Computing*, 8:42, 04 2024.

[13] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[14] Thomas Hoch, Jorge Martinez-Gil, Mario Pichler, Agastya Silvina, Bernhard Heinzl, Bernhard Moser, Dimitris Eleftheriou, Hector Diego Estrada-Lugo, and Maria Chiara Leva. Multi-stakeholder perspective on human-AI collaboration in industry 5.0. In *Artificial Intelligence in Manufacturing: Enabling Intelligent, Flexible and Cost-Effective Production Through AI*, pages 407–421. Springer Nature Switzerland Cham, 2023.

[15] Rosco Hunter, Richard Moulange, Jamie Bernardi, and Merlin Stein. Monitoring human dependence on AI systems with reliance drills. *arXiv preprint arXiv:2409.14055*, 2024.

[16] Ahmed Dawod Mohammed Ibrahum, Manzoor Hussain, and Jang-Eui Hong. Deep learning adversarial attacks and defenses in autonomous vehicles: A systematic literature review from a safety perspective. *Artificial Intelligence Review*, 58:28, 2025.

[17] Alexandra D Kaplan, Theresa T Kessler, J Christopher Brill, and Peter A Hancock. Trust in artificial intelligence: Meta-analytic findings. *Human factors*, 65(2):337–359, 2023.

[18] Uwe Klein, Jana Depping, Laura Wohlfahrt, and Pantaleon Fassbender. Application of artificial intelligence: Risk perception and trust in the work context with different impact levels and task types. *AI and Society*, 39(5):2445–2456, 2024.

[19] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

[20] Zhuoran Lu, Zhuoyan Li, Chun-Wei Chiang, and Ming Yin. Strategic adversarial attacks in AI-assisted decision making to reduce human trust and reliance. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023.

[21] Mahsan Nourani, Joanie King, and Eric Ragan. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 112–121, 2020.

[22] Frank Pajares. Self-efficacy beliefs in academic settings. *Review of educational research*, 66(4):543–578, 1996.

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[24] Christina Pazzanese. Ethical concerns mount as AI takes bigger decision-making role. *Harvard Gazette*, 2020.

[25] Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. Towards making the most of chatgpt for machine translation. In *Findings of EMNLP 2023*, 2023.

[26] Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *International conference on machine learning*, pages 5231–5240. PMLR, 2019.

[27] Akshay Rangesh, Nachiket Deo, Ross Greer, Pujitha Gunaratne, and Mohan M Trivedi. Autonomous vehicles that alert humans to take-over controls: Modeling with real-world data. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 231–236. IEEE, 2021.

[28] Umair Rehman, Farkhund Iqbal, and Muhammad Umair Shah. Exploring differences in ethical decision-making processes between humans and ChatGPT-3 model: A study of trade-offs. *AI and Ethics*, 5(1):279–289, 2025.

[29] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. A missing piece in the puzzle: Considering the role of task complexity in human-AI decision making. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, page 215–227, 2023.

[30] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. Ignore, trust, or negotiate: Understanding clinician acceptance of AI-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

[31] Vijay Veerabadran, Josh Goldman, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, Jonathon Shlens, Jascha Sohl-Dickstein, Michael C Mozer, et al. Subtle adversarial image manipulations influence both human and machine perception. *Nature Communications*, 14(1):4933, 2023.

[32] Maria Virvou, George A Tsihrintzis, and Evangelia-Aikaterini Tsichrintzi. Virtsi: A novel trust dynamics model enhancing artificial intelligence collaboration with human users–insights from a chatgpt evaluation study. *Information Sciences*, 675:120759, 2024.

[33] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. From human-human collaboration to human-AI collaboration: Designing AI systems that can work together with people. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, pages 1–6, 2020.

[34] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 1–12, 2019.

[35] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9):2805–2824, 2019.

[36] Chunpeng Zhai, Santoso Wibowo, and Lily D Li. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments*, 11(1):28, 2024.

[37] Shuning Zhang, Hui Wang, and Xin Yi. Exploring collaboration patterns and strategies in human-AI co-creation through the lens of agency: A scoping review of the top-tier HCI literature. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–43, October 2025.

[38] Yi Zhu, Taotao Wang, Chang Wang, Wei Quan, and Mingwei Tang. Complexity-driven trust dynamics in human–robot interactions: Insights from AI-enhanced collaborative engagements. *Applied Sciences*, 13(24), 2023.