Examining Trust Erosion in Human-AI Collaboration: Unpacking the Impact of Severity and Timing of Misclassifications

ALICIA FREEL*, Indiana University, USA
SABID BIN HABIB PIAS, Indiana University, USA
SELMA SABANOVIC, Indiana University, USA
APU KAPADIA, Indiana University, USA

With the recent advancements in Artificial Intelligence (AI), the issue of user trust in AI has become a crucial aspect of human-AI collaboration. However, the impact of the severity and timing of mistakes in an AI-assisted system, where users rely on AI-generated outcomes, has been understudied. This research aims to investigate whether the loss of user trust in AI systems varies based on the severity of mistakes and their timing. We hypothesize that severe errors in classification will decrease trust in high-stakes scenarios, regardless of when they occur. Conversely, we predict that mistakes in low-severity scenarios will result in greater loss of trust in the AI classifier if they occur towards the end of the interaction compared to the beginning. By instructing algorithms to prioritize accurate identification in situations where a mistake could be severe, developers can optimize the training of their AI models to focus on key metrics that influence trust erosion and enhance the efficacy of human-AI collaboration by deploying human-centered approaches to trust in the design process.

CCS Concepts: • **Human-centered computing**;

ACM Reference Format:
Alicia Freel, Sabid Bin Habib Pias, Selma Sabanovic, and Apu Kapadia. 2024. Examining Trust Erosion in Human-AI Collaboration: Unpacking the Impact of Severity and Timing of Misclassifications. 1, 1 (April 2024), 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

In today's rapidly evolving technological environment, artificial intelligence (AI) has seamlessly integrated into our daily lives. From devices such as smart voice assistants and robotic vacuum cleaners to more advanced technologies such as large language models (LLMs) and self-driving vehicles, AI has rapidly fostered collaboration between humans and AI. This relationship, evident in tasks such as driving and joint problem-solving, has transformed our perception of AI from a mere tool to a valued teammate [5]. This shift highlights the importance of studying the dynamics of human-AI interaction to advocate for a human-centric AI design, along with the integration of elements focused on preventing the loss of trust in the AI development process [2].

Artificial intelligence (AI) developers have historically focused on studying human behavior to create mathematical models and replicate human logic [8]. Recently, there has been a shift toward a more human-centered approach in AI development, emphasizing the importance of considering human requirements, perspectives, and actions during the

Authors' addresses: Alicia Freel, anfreel@iu.edu, Indiana University, P.O. Box 1212, Bloomington, Indiana, USA, 47408; Sabid Bin Habib Pias, Indiana University, Bloomington, USA; Selma Sabanovic, Indiana University, Bloomington, USA, selmas@indiana.edu; Apu Kapadia, Indiana University, Bloomington, USA, kapadia@indiana.edu.

Manuscript submitted to ACM

design process [8, 10]. This change was driven by users who expressed that they did not want to work with AI due to low levels of trust when interacting with it. This led developers to prioritize collaboration and trust as key factors for successful AI integration [8]. In addition, AI developers have historically been faced with the time consuming process of applying data mining and machine learning algorithms. A study showed that data mining and machine learning researchers spend most of their energy on algorithms to construct models [17]. We propose that if developers better understood what types of data most affected trust, they could optimize the training of their AI models to focus on key metrics that influence trust erosion and enhance the efficacy of human-AI collaboration by deploying human-centered approaches to trust in the design process.

However, additional challenges can arise for developers designing AI systems with human-human trust dynamics in mind. Studies show that people perceive errors made by robots (autonomous systems) differently from the same errors made by humans. The study concluded that people tend to lose trust more rapidly in an erroneous decision-making algorithm than in humans who make comparable errors [12]. Consequently, our study further explores current research in trust dynamics between humans and autonomous systems, as trust dynamics in technology seems to be unique from those seen in human-human dynamics.

Previous studies in Human-Robot Interaction (HRI) have explored trust dynamics between humans and autonomous systems [15]. However, there is limited research on how factors such as the severity and timing of misclassifications affect human trust in robots. Some studies have analyzed how the severity of misidentification influences trust in autonomous systems [15], while others have investigated the impact of explanation timing (used by autonomous systems for trust repair) on human trust [15]. However, the relationship between misclassification severity, timing, and trust outcomes in high-stake AI systems remains underexplored.

A study revealed that more severe violations resulted in less effective explanations as trust repair strategies [4]. Meanwhile, Luebbers et al. conducted a study to explore how the effect of serial position (early, midpoint, or late task failures) of mistakes influenced participants' recall of robot competence. The results showed that post-experiment competence and trust ratings were markedly lower in the late task failure condition, indicating a recency bias [13].

Our study focuses on the dynamics of trust when human and artificial intelligence (AI) classifiers work together as a team. We argue that the severity and timing of misclassifications affect trust in high-stakes environments. We hypothesize that highly severe misclassifications reduce trust in high-stake situations, regardless of when they occur. Meanwhile, we predict that less severe misclassifications may only significantly decrease trust if misidentified at the end of the interaction with an AI-classifier, rather than at the beginning. Our research highlights the importance of understanding how the timing and severity of misclassifications influence trust between humans and AI, considering cognitive biases such as recency bias.

In our study, we examine high-stake contexts for AI systems to study the dynamics of trust between severity and timing. In the literature, how humans trust AI classifiers in situations where the stakes are higher is understudied. However, Corriea et al. conducted one such study between humans and robots. They found that if the consequence of the robot's failure was more severe, its trust repair strategies were unable to mitigate the negative impact on perceived trust levels, compared to when the consequences were less severe [4].

In addition, our study focuses on misclassifications due to the common occurrence of false positive errors in high-stake real-world scenarios. False positives occur when non-threats are mistakenly identified as threats, often following a 'better safe than sorry' defense rule in threat detection and cybersecurity intrusion detection systems [6, 9].

High-stakes situations that we consider in our study include those seen in military scenarios, where an AI assists in identifying threats and offers strategic guidance to respond to advanced threats proactively [19]. At the same time,

AI applications contribute to the identification and mitigation of harmful content on social media platforms, working in tandem with human moderators to provide faster and more accurate output [7]. We considered these situations as high-stakes scenarios, due to the severe negative consequences (failure of critical infrastructure, failed search and rescue missions, the censoring of life-saving content on social media) if such threats were ignored or misclassified.

To better understand the effects of the severity and timing of misclassifications on trust in high-stakes uses of AI systems, this study focuses on two central research questions.

RQ1: How does the severity of false positive misclassifications impact human trust in an AI system?

RQ2: How does the timing of false positive misclassifications impact human trust in an AI system?

RQ3: How do the severity and timing of false positive misclassifications jointly impact human trust in an AI system?

We propose that the results of this study could help improve AI systems by refining their design to be more human-oriented and more trustworthy. Our results would also highlight to AI developers the importance of minimizing the impact of severe misclassifications.

2 BACKGROUND AND MOTIVATION

Research on trust in AI has evolved. Initially, AI researchers focused on human behavior to create mathematical models for AI to mimic human logic. Recently, there has been a shift towards a more human-centered approach in AI development, in which user perspectives are considered. This change was driven by users struggling to trust AI, leading developers to prioritize collaboration with users. Trust is now seen as crucial for successful AI integration, especially in scenarios with high autonomy and decision-making abilities [8].

Furthermore, based on previous literature, it is understood across contexts that trust can be conceptualized as a tendency to take a meaningful risk while believing in a high chance of a positive outcome. Trust is also explained as being a dynamic concept that is prone to changes based on the behavior of the trusted agent, along with contextual factors, citing that trust in an agent can increase over time, but can also reduce trust. **Previous research indicates that while trust between humans increases with time through frequent interactions, trust in technology usually decreases due to errors and malfunction** [8].

The cause might be related to the disparity in how humans perceive errors committed by humans versus robots. Research comparing trust erosion in human-human collaboration and human-robot collaboration has revealed that errors made by robots are perceived differently from those made by humans. Studies indicate that individuals tend to see machines as nearly flawless, a concept referred to as automation bias. Consequently, when a machine fails, trust diminishes more rapidly compared to when a human makes an error, as humans are acknowledged to be inherently prone to errors. This is in line with the concept of algorithm aversion, where people tend to lose confidence more quickly in a faulty decision-making algorithm than in humans committing similar errors. Evidently, breaches of trust by machines are evaluated and interpreted distinctively from those by humans [12]. Consequently, the lack of trust that arises from AI errors has been shown to result in a reduced response to AI notifications and disregard for its recommendations. One study cited that an excessive number of alarms, in any system, contributes to alarm desensitization, mistrust, and lack of human responsiveness [1].

In high-stakes scenarios, such as cybersecurity, professionals often face a high volume of false positive alerts in their Security Operations Centers (SOC), necessitating manual review. This verification process is time-consuming and can result in alert fatigue and desensitization. As mentioned above, these excessive alerts can lead to decreased trust, desensitization, and decreased responsiveness [1]. Therefore, there is a growing emphasis on the integration of Artificial Intelligence (AI) in the enhancement of Intrusion Detection Systems (IDS) to improve their ability to detect and classify

network traffic and pinpoint unusual behaviors [14, 16]. By instructing algorithms incorporated into IDS systems to prioritize accurate identification of severe threats from non-threats (false positives), developers can optimize the training of their AI models to focus on key metrics that influence trust erosion, reducing false positives, and increasing human trust as a result.

Human-Centered Interaction Review Previous studies in Human-Centered AI have explored design techniques that influence a person's acceptance of AI. A study by Kocielnik et al. created a study that aimed to understand the impact of different types of error (avoiding false positives versus false negatives). They showed how different types of errors can lead to vastly different subjective perceptions of accuracy and acceptance. They found that user satisfaction an acceptance of a system that makes more False Positive mistakes can be significantly higher than a system optimized for high precision. They hypothesized that the reason users can easily recover from a False Positive is because the wrong words being highlighted by the AI can be ignored, while False Negatives that have no highlighting are due to the cost of recovery from each of these errors [11]. The cost of recovery is relevant in our study- we are hypothesizing that if the cost of recovery is higher (i.e., the severity of the misidentification is higher), that users will be less likely to trust the AI.

Another study by Cai et al. studied the impact of personal characteristics on user trust in conversational reccommender systems. They found that a user's trust propensity and domain knowledge positively influenced the user's trust in conversational recommender systems. To further strengthen our findings, our aim is to account for these factors, by measuring the user's trust propensity, and their domain knowledge [3]. We believe this will add further depth and enrich our findings due to the similarity between our proposed AI classifier and their AI conversational recommendation system.

Zhang et al. also examined how humans trust automated systems in collaborative environments. In their study, they found that participants trusted and relied on the AI the most when it complimented the users weakness (i.e. the AI was an expert in identifying all fake shapes, and the participant was good at only identifying regular shapes. They also found that the increase in trust and reliance in the AI is not due to a higher absolute level of expertise, and found that even when the AI made glaringly obvious mistakes in the complementary condition, participants were still shown to trust and rely on the AI when attempting to identify the regular shapes (which humans might presume to be the easy shapes to identify.) The authors claim that this indicates that subjective trust in an AI does not depend on absolute factors such as competence, predictability, and reliability [20]. This is directly related to our study, as we also intend to promote collaboration between our AI classifier and participants. It is interesting to note that, based on this study's results, the AI can be completely incompetent in one area but still remain trustworthy to participants. Our study would take an interesting twist, by identifying how severity and timing may influence trust in AI. In high-severity circumstances, it would be interesting to see if and how participants trust the AI when there are misclassifications made, and how that impacts human-AI collaboration.

Human-Robot Interaction Literature Review Previous studies on human-robot interaction (HRI) have explored trust dynamics between humans and autonomous systems. However, there is limited research on how factors such as the severity and timing of misclassifications affect human trust in robots. Some studies have analyzed how the severity of misidentification influences trust in autonomous systems, while others have investigated the impact of explanation timing (used by autonomous systems for trust repair) on human trust. However, the relationship between severity, timing, and trust outcomes in high-stake AI systems remains under explored.

One study by Correia et al. revealed that the more severe consequences of violations resulted in less effective explanations as trust repair strategies [4]. Meanwhile, a study by Luebbers et al. conducted a study to explore how the

effect of serial position (early, midpoint, or late task failures) of mistakes influenced the recall of competence. The results showed that the post-experiment competence and trust ratings were notably lower in the late task failure condition, indicating a recency bias [13].

Rossi et al. investigated how human trust depends on the severity and order of errors made by a robot. They suggest that there is a correlation between the severity of the error performed by the robot and the fact that humans do not trust the robot. They found that the trust of the participant was more severely affected when the robot made errors that had severe consequences, and that the participants were less likely to trust the robot when severe errors occurred at the beginning of the interaction [18].

In their research, Correia et al. conducted an experiment using a 2x2 design (including a control group) with a relatively small sample size of 97 participants. The aim was to explore the impact of technical malfunctions in autonomous social robots on trust within collaborative environments. The study revealed that factors such as the type of task, the nature of the error, and the severity of the error can influence how individuals view robots after an error occurs. Their findings suggested that when a social robot exhibits faulty behavior during a joint task, such as a puzzle activity, it tends to be perceived as less reliable. However, the main result indicates that justifying the failure as a recovery strategy can mitigate its negative impact on trust, but only when the consequence of the failure is less severe. On the other hand, **when the failure is more severe, the recovery strategy had no effect on trust.** The researchers proposed that this could be attributed to participants who have higher expectations about the effectiveness of the recovery strategy, with simple justification considered insufficient [4].

Luebbers et al. conducted a human subject research with 53 participants to investigate how the serial position effect affects the recall of competence. The participants watched videos of a robot completing tasks at a consistent level of competence, but the order of the tasks varied depending on the experimental condition (early task failures, midpoint task failures, late task failures). They were asked to rate the robot's competence after each video and at the conclusion of the experiment. They found that while the average rating between videos of robot competence remained stable across conditions, the recalled, post-experiment ratings of competence and trust were significantly lower in the condition with late task failures than in either of the other two conditions, suggesting a notable recency bias [13].

The additional literature on timing provides mixed results on the impact of timing on trust. Gilkson et al. referenced a study by Desai et al. that indicated that initial errors made at the beginning of an interaction had less of a negative impact on trust, compared to errors made in the middle or conclusion of an interaction. In contrast, the authors mention Rossi et al., who discovered that significant errors made by a robot toward the end of an interaction had less of an impact on trust compared to similar errors made at the beginning of an interaction [8].

In line with Luebbers et al.'s findings, we argue that the timing of misclassifications should also be further investigated due to consequential biases such as the recency bias (information presented last among a grouping being the most salient in memory formation) that could in turn affect trust levels in AI-classification systems. Additionally, Luebbers's study had a relatively small number of participants. Our aim is to replicate these results in a larger sample pool to understand how misclassification timing affects trust in AI-classifiers.

Recognizing the limited research, small sample sizes, and the need for context-specific studies on the influence of severity and timing on human trust in high-stakes AI systems, our goal is to conduct a study that can lay the groundwork for future research. The objective is to refine, comprehend, and increase the trust dynamics between humans and AI to facilitate a more reliable and efficient collaboration. Additionally, we aspire to enhance and expedite AI algorithms in practical applications in areas like cybersecurity, national defense, and social media monitoring.

3 PROPOSED METHOD

Our study would explore two specific moderators of trust in a high-stakes AI classification system: the severity and timing of the impact of misclassifications on trust. The main study will be an online survey and include seven conditions in a design between subjects. Participants would be randomly selected for one of seven conditions:

(1) Control Group: No images are misclassified, and at the end participants are asked to rate their trust in the AI Operator.
(2) High-severity images are misclassified at the beginning of the survey, and then participants are asked to rate their trust in the AI Operator.
(3) High-severity images are misclassified at the end of the survey, and then participants are asked to rate their trust in the AI Operator.
(4) High-severity images are misclassified randomly, and then participants are asked to rate their trust in the AI Operator.
(5) Low-severity images are misclassified at the beginning of the survey, and then participants are asked to rate their trust in the AI Operator.
(6) Low-severity images are misclassified at the end of the survey, and then participants are asked to rate their trust in the AI Operator.
(7) Low-severity images are misclassified randomly, and then participants are asked to rate their trust in the AI Operator.

Within the online survey, participants, regardless of the condition, would receive both prompts based on two scenarios: a scenario where participants are a military operator tasked with identifying potential false positive misclassifications (high-stakes environment) and a social media moderator for children under the age of 17 tasked with identifying false positive images rated by the classifier (low-stakes environment). Quantitative and qualitative information will be collected from participants.

Participants would be presented with an image based on the scenario and asked whether they agree with the AI's classification of the image. An example of a low-severity misclassification made by the AI would be "This is an image of a frozen lake" when, in reality, the image is of a baseball field. This example would simulate an incorrect assessment of the geographic mapping of nearby territory within a defensive military context. If the AI classifier is wrong, there is no real consequence for the safety or livelihood of others. In contrast, a highly severe mistake would be the AI mistaking a cruise ship for an enemy military ship, which could result in casualty if defensive measures are taken.

An example of a low-severity misclassification within the social media context would be a post that shows a joke being deleted by the AI, while a high-severity misclassification would be the deletion of a lifesaving informational post, such as how to administer an epi pen, which, if blocked, could potentially result in lifesaving information not being communicated.

If the AI is indicated as being wrong by the participant, then the participant would be asked to score the severity of the misclassification on a 5-point scale (not at all severe - majorly severe) and asked to explain why they rated the misclassification as the indicated severity.

At the end of the survey, participants would then be asked to complete a trust scale measurement to understand the level of trust in the AI classifier, relative to one of the seven conditions to which they were assigned, concluding the survey. Additional questions would also be asked surrounding the user's propensity to trust others and their expertise in AI, to account for impacting factors found in previous studies.

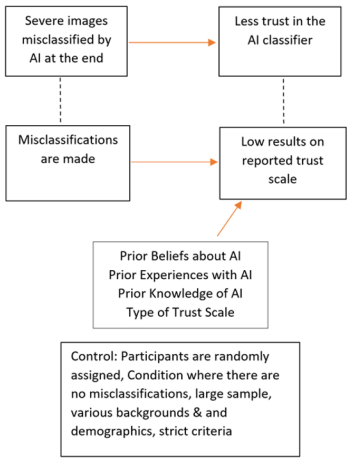

Fig. 1. Box and Arrow Diagram depicting our hypothesis: when misclassifications are made, the result is low reported levels of trust; when severe misclassifications are made at the end of the study, this will result in even less trust in the classifier. Additionally, we aim to capture how prior beliefs, experiences, knowledge, and type of trust scale can also impact the reported levels of trust in the AI system

3.1 Stimulus Validation

Before the main survey is launched, the study would involve collecting and validating a set of false positive misidentifications stimulus for image classification scenarios in military and social media contexts. Participants would rate image severity on a 5-point scale (not at all severe to majorly severe) during the norming process, ensuring the relevance and appropriateness of the sample size and stimulus set. The stimulus validation process will play a crucial role in ensuring the relevance and appropriateness of the image set used in this study.

4 DISCUSSION

Recent advances in artificial intelligence have paved the way for stronger collaboration between humans and AI. However, there is limited research on how factors such as the severity and timing of misclassifications affect trust in high-stake situations. Past research has demonstrated that AI system errors can decrease trust and impact people's willingness to follow AI advice [12]. It is essential to understand how trust decreases between humans and AI systems to promote successful collaboration in real-world scenarios. Our study aims to fill the research gaps presented in existing literature by examining the effect of severity and timing on trust dynamics when an AI classification partner makes mistakes in high-risk situations. We propose that highly severe misclassifications will reduce trust in high-stake situations, regardless of when they occur. In contrast, we predict that less severe misclassifications will lead to an overall loss of trust in the AI classifier only if misclassifications occur at the end of the interaction, rather than at the beginning. Furthermore, by investigating cognitive biases that may impact trust relative to the timing of misclassifications, such as recency bias, we aim to provide more context to the importance of the timing and severity of misclassifications impact on trust. Real-world applications of these findings would guide the refinement of AI design and algorithms. By instructing algorithms to prioritize accurate identification in situations where an error could be severe, developers could streamline the training of their AI to focus on critical metrics that influence trust erosion. These design improvements

would consequently promote greater collaboration between humans and AI in high-stake environments by considering human-centered approaches to trust in the design process.

4.1 Impact of Severity and Timing

Severity's Role in Loss of Trust. The effect between severity and loss of trust, especially in the face of severe misclassifications, would accentuate the need for AI systems to prioritize accuracy, particularly in high-stakes environments. The study would reaffirm that serious errors have a substantial and lasting impact on trust, highlighting the importance of minimizing such errors in critical applications such as military operations and the moderation of social media content.

Timing's Influence on Trust Repair. The "recency bias" effect shown in previous research and then also considered in our own research would underscore the significance of the timing of misclassifications in trust if less severe misclassifications shown at the end result in greater trust lost compared to those same misclassifications shown at the beginning. A recency bias would also provide context as to why this effect is observed, providing an improved working framework for AI developers.

4.2 Implications for Human-AI Collaboration:

In the domain of cybersecurity and national security, the issue of false-positive threats is often highlighted as a major concern [1]. Neglecting such information can pose significant risks, which makes the outcomes we achieve highly consequential. Fostering trust between human operators and AI systems is crucial to successful identification and human response to threats. This research would offer practical recommendations for these critical domains, emphasizing the need for AI systems not only to prioritize accuracy but also to demonstrate consistent performance over time, while considering the intricate social dynamics involved in human-AI collaboration.

Acknowledging the scarcity of research, the small scale of samples, and the necessity for studies tailored to specific contexts regarding the impact of severity and timing on human trust in high-stakes AI systems, our aim is to carry out a study that can establish a foundation for subsequent research. The aim is to improve the understanding of the dynamics of trust between humans and AI to promote a more dependable and effective partnership. Furthermore, we aim to improve and accelerate AI algorithms in real-world scenarios for use in fields like cybersecurity, national defense, and social media moderation.

5 CONCLUSION AND FUTURE DIRECTION

In conclusion, the primary goal of our study is to advance the comprehension of trust dynamics in Human-AI collaborative systems. Through the examination of the intricate impacts of severity and timing on trust, we aim to enhance the current body of knowledge and address critical knowledge gaps. Our research focuses on investigating the influence of misclassifications in high-stakes AI systems on user trust, specifically analyzing whether the severity of the error and the timing of its occurrence during the user interaction affect trust levels differently. We posit that errors in high-severity scenarios will uniformly reduce trust in these crucial systems, regardless of when the mistake occurs. Conversely, in less severe scenarios, we predict that trust in the AI system will be more significantly undermined if errors occur towards the end of the user interaction compared to those at the beginning. The implications of our findings go beyond theoretical insights, offering practical benefits for the advancement and deployment of AI systems, particularly in critical domains like cybersecurity, national security, and social media moderation. Future research

endeavors could explore various trust restoration strategies based on the findings of this study. Moreover, it is essential to prioritize real-world experimentation and dynamic trust frameworks to ensure the successful integration of AI into decision-making processes. Ultimately, our study aims to lay a robust groundwork for further exploration, with the goal of refining and enhancing trust dynamics between humans and AI classifiers to facilitate more dependable and efficient Human-AI collaboration.

ACKNOWLEDGMENTS

This material is based upon work supported by the Department of Defense via Purdue University under funding agency 13000844-031.

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] Carolina Centeio Jorge Siddharth Mehrotra Anna-Sophie Ulfert, Eleni Georganta and Myrthe Tielman. 2023. Shaping a multidisciplinary understanding of team trust in human-AI teams: a theoretical framework. *European Journal of Work and Organizational Psychology* 0, 0 (2023), 1–14. https://doi.org/10.1080/1359432X.2023.2200172

[3] 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. https://doi.org/10.1145/3491102.3517471

[4] Filipa Correia, Carla Guerra, Samuel Mascarenhas, Francisco Melo, and Ana Paiva. 2018. Exploring the Impact of Fault Justification in Human-Robot Trust.

[5] Christopher Flathmann, Beau G. Schelble, Patrick J. Rosopa, Nathan J. McNeese, Rohit Mallick, and Kapil Chalil Madathil. 2023. Examining the impact of varying levels of AI teammate influence on human-AI teams. *International Journal of Human-Computer Studies* 177 (2023), 103061. https://doi.org/10.1016/j.ijhcs.2023.103061

[6] Paul Gilbert. 1998. The evolved basis and adaptive functions of cognitive distortions. *British Journal of Medical Psychology* 71, 4 (1998), 447–463. https://doi.org/10.1111/j.2044-8341.1998.tb01002.x arXiv:https://bpspsychub.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-8341.1998.tb01002.x

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

[8] Ella Glikson and Anita Woolley. [n. d.]. Human Trust in Artificial Intelligence Review. ([n. d.]).

[9] 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. https://doi.org/10.1109/MCOM.2012.6163595

[10] Alejandro Jaimes, Daniel Gatica-Perez, Nicu Sebe, and Thomas S. Huang. 2007. Guest Editors' Introduction: Human-Centered Computing–Toward a Human Revolution. *Computer* 40, 5 (2007), 30–34. https://doi.org/10.1109/MC.2007.169

[11] 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. https://doi.org/10.1145/3290605.3300641

[12] 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.

[13] Matthew B Luebbers, Aaquib Tabrez, Kanaka Samagna Talanki, and Bradley Hayes. [n. d.]. Recency Bias in Task Performance History Affects Perceptions of Robot Competence and Trustworthiness. ([n. d.]).

[14] 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, 3 (2023), 30–37. https://doi.org/doi:10.2478/kbo-2023-0072

[15] 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.

[16] Benoit Morel. 2011. Artificial intelligence and the future of cybersecurity. In *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence* (Chicago, Illinois, USA) *(AISec '11)*. Association for Computing Machinery, New York, NY, USA, 93–98. https://doi.org/10.1145/2046684.2046699

[17] M. Arthur Munson. 2012. A study on the importance of and time spent on different modeling steps. *SIGKDD Explor. Newsl.* 13, 2 (may 2012), 65–71. https://doi.org/10.1145/2207243.2207253

[18] 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.

[19] Clay Wilson. 2020. *Artificial Intelligence and Warfare.* Springer International Publishing, Cham, 125–140. https://doi.org/10.1007/978-3-030-28285-1_7

[20] 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. https://doi.org/10.1145/3491102.3517791

Received 23 February 2024