The impact of AI literacy on the emotional reaction to AI mistakes in human-AI collaboration

Beatrice Beretta

Faculty of Behavioural, Management & Social Sciences, University of Twente

Module 12, 202000384: BSc Psychology of Conflict, Risk and Safety

Esther S. Kox

January 24, 2023

Abstract

The growth of artificial intelligence (AI) has had a significant impact on human life, offering new opportunities for collaboration. However, integrating AI into various aspects of society has also presented challenges, for example mistakes made by AI systems. This research investigates the relationship between individuals' understanding of AI, their trust and attitude towards AI, and their emotional responses to Ai mistakes. The study suggests that having an appropriate level of AI literacy is crucial in shaping how people perceive AI. To explore emotional responses to AI mistakes, the study uses different real-life scenarios and investigates how levels of AI literacy affect trust and attitude towards AI. The study proposes that higher levels of AI literacy would lead to reduced emotional distress after the exposure to AI mistakes. The study also aims to demonstrate how the level of AI literacy is associated with people's trust and attitudes towards AI. Moreover, the research investigates how participants' trust and attitude towards AI might change after exposure to AI mistakes. 185 participants took part in the study, with their level of AI literacy assessed using the Meta AI Literacy Scale (MAILS), trust measured using the Human-Computer Trust model, and attitude evaluated based on Hidalgo's model. The results do not provide with significant results in regard to the relation between AI literacy and emotional response to AI mistakes. On the other hand, the study shows how the level of AI literacy affects people's trust and attitude towards AI. Finally, the study shows that there is a significant decrease in trust and attitude towards AI after exposure to mistakes made by AI systems.

Keywords: psychology of risk & safety, AI, trust & attitude towards AI, AI literacy, AI-driven mistakes, emotional response to AI failures

Introduction

The field of artificial intelligence (AI) has experienced significant growth in recent years (Wang et al., 2020). This technological advancement has opened numerous new opportunities for collaboration between humans and AI (Rossi, 2021). Its application ranges from virtual assistants (e.g., Siri, Alexa), which functions as a personal assistance tool in our everyday life easier, to self-driving cars, such as Tesla Autopilot (Dikmen & Burns, 2016) and Waymo, formerly Google's Self-Driving Car Project (Rosenband, 2017). Thus, AI has become a crucial aspect of our lives (Elliott, 2019). Its efficiency has transformed not only the way we live our daily lives but also bigger industries such as healthcare, finance, and education.

Blinded by the many advantages of AI use and fascination with the new technologies and the fact that AI is becoming more common, people tend to show an overall positive attitude and trust towards AI (Choung et al., 2022). However, human-AI collaboration can also present important challenges (Wang et al., 2020). One of these challenges is the occurrence of different types of mistakes in AI systems. Such mistakes are likely to negatively affect people's overall trust and attitude towards AI. To overcome these issues, previous research already explored multiple strategies (e.g. apologies) to restore trust in, and a positive attitude towards AI (Kox et al., 2021, 2022). A practical application of these restoring strategies is, for example, shown by ChatGPT-3.5, when the chatbot apologises for miscomprehending the user’s request and provides an alternative response. The expression of regret provided by the AI helps to rebuild trust (Gillath et al., 2021) but not for the human user to understand the reason for the AI mistake. However, research in the field of AI has yet to thoroughly examine the intricate emotional and psychological components that influence individuals' reactions to errors made by artificial intelligence.

Research suggests that understanding how people comprehend AI plays a vital role in determining the extent to which they trust and perceive AI systems (Gillespie et al., 2023). The capacity of individuals to comprehend, engage with, and make informed determinations about AI technologies is referred to as ‘AI literacy’ (Ng et al., 2021). AI literacy can also be defined as the ability to understand fundamental AI concepts, recognise everyday applications of AI, and evaluate their impact on society (Long & Magerko, 2020). AI literacy enables individuals to make informed decisions, engage in discussions about AI's role in society, and adapt to advancements driven by AI (Wagner, 2021). In today's world, where AI is integrated into many industries and aspects of our lives, gaining a comprehensive understanding of how people emotionally respond to mistakes made by AI is a crucial area of investigation. Learning whether higher levels of AI literacy foster more constructive and well-informed responses during such instances would provide valuable insights for further exploration.

This research explores the relationship between the level of an individual’s AI literacy, their trust in and attitude towards AI and the experience of AI-driven mistakes. First, some background information will be given on AI and the trust and attitude towards AI in human-AI collaboration. Second, information will be given on the concept of AI-driven mistakes and which type of scenarios will be used in the current study. Last, the concept of AI literacy and expectations of interactions with the perception of AI-driven mistakes will be explored.

Concludingly, the primary aim of this study is to investigate how AI literacy influences the perception of AI mistakes by humans with the following research question “To what extent does the level of AI literacy influence the emotional responses and perception of AI mistakes in human-AI collaboration?”.

Trust and attitude towards AI

In the domain of human-AI collaboration, there are two factors that impact the success and acceptance of human-AI collaboration. Trust is the first factor that plays a fundamental role in determining how individuals interact with AI technologies (Lukyanenko et al., 2022). Additionally, the attitude that individuals present towards AI is a necessary factor for effective partnerships between humans and AI systems in the various fields of human-AI collaboration, such as business or healthcare (Shin, 2021).

It is expected that as time passes and technology continues to advance, attitudes towards AI may shift and trust may grow if these systems demonstrate their utility and fairness in everyday life (Rossi, 2021). However, it is important to note that public sentiment towards AI remains dynamic. The study from Lee and See (2004) shows the importance of establishing an optimal degree of confidence and faith in AI systems by considering 'calibration trust' as a way in which humans can assess both the capabilities and limitations of AI systems to engage in suitable reliance and productive cooperation with them. More specifically, if a person witnesses reduced trustworthiness (e.g. any AI mistake), their trust in the system should diminish. An example of poor trust calibration can be found in the study by Buçinca et al. (2021). The study shows that people tend to overly rely on AI systems, even after repeated mistakes. The same research aims to assess the level of overreliance by exposing the participants to AI mistakes and providing an explanation for the mistake. Even after the explanation, the overreliance still exists and tends to increase. Nevertheless, the majority of people display a high level of trust in AI and hold a positive attitude towards it. These individuals view AI as a valuable tool that has the potential to enhance efficiency and convenience in various aspects of life (Rossi, 2021).

AI mistakes

With the growing integration of AI systems in our daily life, it is crucial to recognize that these systems are fallible and should not be perceived as infallible entities. This ensures a realistic perspective on human-AI collaboration, that promotes awareness on ethical considerations and supports the development of AI systems that align with societal values and needs. AI is susceptible to errors of various nature. A first example is the reliance of AI systems on the data they are trained on. Incomplete, biased, or inaccurately representative training data can lead to errors in AI performance (Raji et al., 2022). Next, AI models may struggle to fully comprehend context and nuances, lacking the flexibility to adapt, especially in unfamiliar situations. AI systems do not possess human understanding and consciousness. The lack of this characteristic may generate content misaligned with societal norms or lack a sense of appropriateness (Yampolskiy, 2016). Furthermore, simple technical issues (e.g. software bugs or hardware failures) can contribute to errors in AI execution (Yampolskiy & Spellchecker, 2016). Finally, AI lacks a comprehensive understanding of ethical and legal intricacies (Hristov, 2016) which makes it susceptible to unintentional acts of plagiarism or copyright infringement.

Emotional response to AI mistakes and failure

Understanding how people react to mistakes made by AI systems involves examining the various emotions and responses that users may undergo during their interactions with these systems. The emotional reactions that individuals may experience could encompass feelings of frustration, confusion, annoyance, disappointment, weakened trust, or even amusement in certain instances (Kocielnik et al., 2019). These reactions can manifest themselves differently among users.

People's emotional responses to AI mistakes can vary based on the nature and impact of the mistake, as well as their personal experiences and attitudes towards technology. Some common emotional responses can include feelings of frustration, annoyance or anger, especially in situations where accuracy is crucial. People's emotional reactions – which can range from frustration to disappointment – play a significant role in determining whether they will continue using the technology (Shank et al., 2019).

Previous research examined trust violations leading to a more negative affect towards AI (Alarcon et al., 2023). The current study aims to add to the current scientific community by investigating the emotional response to AI mistakes in relation to the level of AI literacy.

Purpose of the current study and Hypotheses

This study uses diverse real-life scenarios demonstrating AI mistakes to examine the impact that AI fallibility has on human-AI interaction. The current study aims to analyse various relationships using three main constructs. Namely, the concept of trust and attitude and how these levels vary depending on an individual's AI literacy. Additionally, the study also explores the concept of AI literacy and how this influences people's emotional responses to different AI mistakes.

The first hypothesis (H1) predicts that individuals with higher AI literacy will exhibit lower emotional distress when presented with AI mistake scenarios compared to those with lower AI literacy.

Based on the fact that people tend to approve AI even when they do not possess a high level of these systems (Scantamburlo et al., 2023), the second hypothesis (H2) posits that individuals with low AI literacy will demonstrate trust in AI, despite their limited understanding.

In relation to the second hypothesis, the third hypothesis (H3) suggests that participants with low AI literacy will exhibit a positive attitude towards AI.

Finally, the study will explore whether participants' trust in AI remains stable after exposure to AI mistake scenarios and if there is no significant alteration in attitudes toward AI following the AI mistakes. According to existent literature, individuals tend to maintain trust in AI and sustain positive attitudes despite repeated mistakes and hence, the levels of trust or attitude should remain unchanged after the exposure to AI mistakes (Buçinca et al., 2021; Rossi, 2021).

Method

Design

This study uses a non-experimental design, specifically a correlational approach to investigate the associations between different variables through an online survey. The evaluation included participants' AI literacy, Trust in AI Chatbots, Attitudes toward AI and their Perceptions of AI mistakes. Trust in AI Chatbots and Attitudes toward AI were measured before and after exposure to a series of AI mistake scenarios.

The sample, comprising of 185 participants, exhibited an average age of $M = 33.66$, with the minimum age being 18 and the maximal being 65 with a standard deviation of $SD = 13.51$. The responses of the participants that indicated non-familiarity with AI chatbots, were excluded from the study. A diverse group was gathered through voluntary sampling techniques by utilizing online platforms. In particular, the questionnaire was posted on Survey Circle, an online tool for distributing surveys and gathering responses from different users. Additionally, the survey was also published on the SONA system, a specific online platform that assigns credits for students of Behavioural Sciences at the University of Twente. The survey was accessible to anyone who possessed the survey link, which ensured a wide and varied pool of participants. To ensure data quality and foster a comprehensive understanding of the survey content, only English-proficient participants aged 18 and above were eligible to complete the questionnaire.

The distribution of the educational background of the participants can be found in Table 1 below, which shows that most of the participants who completed the survey have a technical background.

Table 1

Frequencies of participant's educational background

	Distributions
Engineering	39%
Other	28%
Psychology	22%
Communication Sciences	8%
Medicine	2%
Law	1%

Materials

The survey tool utilized in this research was created and executed using Qualtrics, with the following publication on Survey Circle and the SONA system.

An online questionnaire format was employed for the survey using Qualtrics. The demographic data included age, educational background, familiarity with Chatbots and frequency of usage of Chatbots.

AI Literacy was measured using the Meta AI Literacy Scale (MAILS) (Carolus et al., 2023). The MAILS is a comprehensive tool designed to measure individuals' understanding, knowledge, and critical thinking skills related to AI and its applications. The questionnaire consists of 72 items, made of 11-point Likert-scale questions. Participants were asked, for example, to rate on a scale from 0 to 10 their ability to use AI, with a statement as *"I can operate AI applications in everyday life".* The MAILS uses a combination of different scales that each measure a different aspect of AI literacy (e.g. AI literacy apply, AI literacy ethics). Overall, the questionnaire showed good internal consistency and high reliability (all $\alpha > .81$). For this study,

not all items from the original questionnaire were used. More specifically, only items referring to "Use & Apply AI", "Know & Understanding AI" and "AI Ethics were used for this study.

Trust in AI chatbots was measured using The Human-Computer Trust model (HCTM) established by Gulati et al. (2019) to assess individuals' perceptions of trust in technology and specifically in AI. The questionnaire included 12 items, to which participants responded on a 5-point Likert-scale (from 1 = "strongly disagree" to 5 = "strongly agree"). For example, participants were asked to rate on a scale from one to five how much they agreed or disagreed with the following statement *"It is risky to interact with AI chatbots".* The scale showed good internal consistency reliability ($\alpha > .85$).

Attitude towards AI was measured using the work of Hidalgo et al. (2021). The participants were asked to complete a 3-item questionnaire (see Appendix A), on a 5-point Likert scale (from 1 = "no" to 5 = "yes, very"). Participants were asked to rate from one to five how much they disagree or agree with a statement such as, *"AI makes me feel worried".* Hidalgo's original publication provided a comprehensive overview of the model's development and validity (Hidalgo et al., 2021). Additionally, Hidalgo's model has been previously approved and used within the scientific community (Ho et al., 2023). The items of the scale showed acceptable internal consistency reliability ($\alpha > .70$).

For the screening and analysis of the data, Excel and R Studio were used.

Procedure

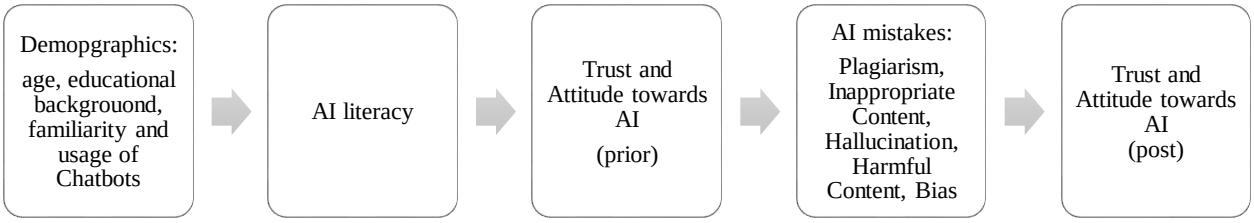
Before participating in the survey, each participant was presented with an informed consent form outlining the study's risks, the right to retreat at any point from the survey and the confidentiality and anonymity of the collected responses. The aim of the study was not included in the consent form to avoid any form of bias from the participants. Each participant provided

electronic consent by clicking a checkbox before proceeding with the survey. Participants accessed the survey on Qualtrics through a URL provided by the researcher.

The survey consisted of five sections (see Figure 1). First, demographic data were collected from the participants. The second section measured the AI Literacy level. Following that, the survey integrated an assessment of respondents' degrees of Trust and Attitude towards AI, before the AI mistakes. After the preliminary assessment of Trust and Attitude, participants were provided with a range of scenarios that portrayed mistakes made by AI. The scenario depicted an example of Plagiarism, an example of the creation of Inappropriate Content from an AI system, an example of Hallucination, one depicted Harmful Content and one showed Bias (see Appendix B for scenarios). For each scenario, participants' Emotional Responses to these mistakes were examined using a series of questions evaluating emotional distress (see Appendix B, Emotional Response), aimed to measure the emotional response to each of the five types of AI mistakes. Finally, the last section related to the Attitude and Trust towards AI after exposure to the AI mistakes.

Figure 1

Study design



Data Analysis

Firstly, descriptive statistics were calculated to provide a summary of the main X_1 characteristics of the study variables. Measures as means and standard deviations were computed

for demographic data and all the variables. For AI literacy, an average between the various subsections of the META questionnaire was computed. For Trust and Attitude towards AI, one average was computed for the items related to prior measurements and one separate average was computed for the post-measurements. Another average was computed to measure the overall level of Emotional Response from the questions following the AI mistakes. Next, the variables were checked for parametric assumptions, specifically normality with the Shapiro-Wilk Test and boxplots and homogeneity of variance with Bartlett's Test.

To address the research questions and the hypotheses, inferential statistical analyses were conducted and for all the hypotheses, the alpha level was set at .05 to test for the statistical significance of the results. To test the first hypothesis, Spearman's Rank Order analysis was conducted to understand the relationship between the level of AI Literacy and the Emotional Response to AI mistakes. For the second hypothesis, another Spearman's Rank Order analysis was conducted to understand the relationship between the level of AI Literacy and the participant's levels of Trust and Attitude towards AI before exposure to AI mistakes. To address the third hypothesis, a paired sample t-test was used to examine any potential change in the initial levels of Trust and Attitude towards AI after exposure to the AI mistakes scenarios.

Results

Manipulation check and assumptions

Of the sample, 94% are familiar with Chatbots, e.g. ChatGPT. The responses from the remaining 6% of the participants were excluded from the analysis. Regarding the usage of Chatbots, e.g. ChatGPT, most participants (36%) declared to make use of such AI systems at once a week. For the other frequency of usage of Chatbots, see Table 3 below.

Table 2

Usage of Chatbots, e.g. ChatGPT

	Distributions
Once a week	36%
Once a month	23%
Almost every day	18%
Once a year	13%
Never	10%

To meet the assumptions for parametric analysis the data were tested for normality and equality of variance. For AI literacy level (M = 5.62; SD = 1.88), the Shapiro-Wilk test was performed to test for normality and showed evidence of normality (W = .98, p = .05). However, after visual examination of the boxplots, it was possible to conclude that the assumption of normality was not supported.

For the variable AI mistakes (M = 3.33; SD = .68), the Shapiro-Wilk test was performed to test for normality and showed evidence of non-normality (W = .98, p = .03). The variables indicating the level of trust towards AI before the mistakes (M = 2.74; SD = 1.03) and after the exposure to the AI mistakes (M = 2.54; SD = 1.4), both indicated evidence for normality (W =

.92, p = .33; W = .89, p = .12). For the variables indicating the level of attitude towards AI before the mistakes (M = 3.67; SD = 1.10) and after the exposure to the AI mistakes (M = 3.76; SD = 1.14), both indicated evidence normality (W = .99, p = .84; W = .79, p = .09).

The homogeneity of variance assumption was tested using Bartlett's test, which indicated that the assumption had been violated for all the variables. For the AI literacy level, the test revealed $\chi^2(1)$ = 10.78, p = .01, making it possible to reject the assumption of homogeneity of variances. For the second variable, emotional response to the AI mistakes, the test revealed $\chi^2(1)$ = 12.88, p = .01, making it possible to reject the assumption of homogeneity of variances. For the third variable indicating trust towards AI before and after the mistakes are shown, the assumption of homogeneity of variances was rejected at the 0.05 significance level (p <0.05), indicating that the variances across groups are likely to be unequal. The same measurements were found for the variable indicating attitude before and after the scenarios of the AI mistakes were shown.

Hypothesis testing

To test whether participants with higher levels of AI literacy would exhibit less emotional distress when exposed to various AI mistake scenarios (H1), a Spearman's rank correlation analysis was performed. The analysis revealed a non-significant negative correlation (ρ = -.05, p = .49). This means that there is no relation between one’s AI literacy and their emotional distress in response to AI mistake scenarios. Hence, the first hypothesis is rejected.

A Spearman's rank-order correlation was also conducted to examine the relationship between trust towards AI among participants before exposure to AI mistakes and their level of AI literacy. More specifically, if participants’ trust towards AI is influenced by their level of AI literacy. The analysis revealed a significant positive correlation (ρ = 0.08, p = .04), indicating

that there is a positive association between individuals' AI literacy levels and their reported trust towards AI. This means that as the level of AI literacy increases, the reported trust in AI also increases. Hence, the second hypothesis (H2) is not accepted.

For the second hypothesis, another Spearman's rank-order correlation was conducted to examine the relationship between attitude towards AI among participants before exposure to AI mistakes and their level of AI literacy. More specifically, if participants' positive attitude towards AI is negatively influenced by their level of AI literacy. The analysis revealed a statistically significant negative correlation (ρ = -0.09, p = .04) suggesting that as the level of AI literacy increases, the positive attitude towards AI tends to decrease, and vice versa. Hence, the third hypothesis (H3) is accepted.

Two paired-sample t-tests were conducted to assess the possible change of trust and attitude towards AI after exposure to AI mistakes. The analysis revealed a statistically significant difference ($t(1823) = 8.72$, $p < 0.05$). This suggests that there is a significant change in trust towards AI after experiencing AI mistakes. The positive t-value indicates that, on average, trust towards AI increases after the mistakes were shown. Another paired-sample t-test was conducted to assess the possible change in attitude towards AI before and after the occurrence of AI mistakes. The analysis revealed a significant difference ($t(455) = 5.94$, $p = < 0.05$), suggesting that after the exposure to the AI mistakes, people tend to report a positive attitude towards AI. Hence, it is possible to conclude that, in this study, participants' trust and attitude towards AI tends to increase after exposure to AI mistakes.

Discussion

This investigation centred on the question “To what extent does the level of AI literacy influence the emotional responses and perception of AI-driven mistakes in human-AI collaboration?”. Through an examination of emotional reactions and beliefs across different levels of AI literacy, this research’s goal is to provide more insightful knowledge about the dynamics of trust, attitude, and emotional adaptability in the interactions between humans and AI.

Main Findings

The focus of this research was to examine the complex connection between people's understanding of AI (i.e., AI literacy), their emotional reactions to AI mistakes, and their trust and opinions about AI. Participants with higher AI literacy were expected to experience less emotional distress when presented with different AI mistake scenarios, according to the initial hypothesis. Contrary to the initial expectations, the analysis indicated that there is no association between AI literacy and emotional distress. The lack of a link between AI literacy and emotional distress in different AI error scenarios might have been caused by several factors that influence how individuals respond emotionally in this specific context. Psychological factors like emotional intelligence and coping strategies, as well as contextual factors such as familiarity with AI situations and the perceived realism of scenarios, may have played a role in shaping participants' reactions. Various individual differences like personality traits, prior experiences with AI, and sample characteristics like educational background and age could also contribute to the range of emotional responses observed. For example, the fact that the majority of the participants possess a technical educational background might have caused them to not experience emotional distress as they might be more aware of the fallibility of AI systems.

Next, it was expected that participants who had a greater understanding of AI would have low trust and show a more neutral or negative attitude towards AI. The first analysis showed that, in line with the second hypothesis, as AI literacy decreased, the trust towards AI increased and vice versa. This outcome is in line with the study by Scantamburlo et al. (2023) which supports the fact that people tend to present a low level of AI literacy and yet approve of AI. Despite having limited knowledge of AI, people often support these systems due to, for example, fascination and newness associated with these technologies, positive depictions in the media, and an overall optimism about their potential advantages. Furthermore, a preference for simplicity in thinking, and social influence might also have an impact on shaping overall trust towards AI.

However, when investigating the relationship AI literacy and attitude towards AI, the results shows how if the level of AI literacy increases, so does the positive attitude towards AI. This finding does not meet the initial believes. Psychological elements including existing beliefs or perceptions of AI, might have impacted participants' attitudes towards AI regardless of their AI literacy level (De Sá Siqueira et al., 2023). For example, someone who already has pre-existing positive beliefs about the advantages of AI systems may continue to have a positive view of AI, no matter how much they know about AI, so their level of AI literacy. Conversely, someone with negative preconceived notions might maintain a sceptical or negative attitude towards AI, even if they are well-informed about AI systems. These pre-existing thoughts represent ways in which people interpret new information and they can greatly impact their reactions and opinions about AI, regardless of their actual comprehension of AI concepts.

Finally, the study aimed to observe if there would be any notable alteration in trust and attitude towards AI following exposure to various AI mistakes. In line with the existing literature (Buçinca et al., 2021; Rossi, 2021), both the levels of trust and attitude towards AI increased

after exposure to AI mistakes. This emphasizes the tendency for individuals to over-rely on AI systems despite repeated mistakes. People might excessively depend on AI systems even when they make frequent errors because of cognitive biases such as automation bias. This bias causes individuals to believe that technology is flawless (Strauß, 2021). The sunk-cost fallacy can also play a role, as people resist changing their attitudes to rationalize the time and effort they have invested in learning and adopting AI and, as a result, to downplay the significance of AI mistakes (Balakrishnan et al., 2021). Although this finding is in line with the existing literature, it goes against common sense. More specifically, people tend to lower their trust towards something when it makes mistakes, it is in human nature to introduce doubt and scepticism after trust is broken (Schepman & Rodway, 2022). Nevertheless, in the case of mistakes by AI, this study shows how trust does not decrease.

In regard to the demographic data, the age range of participants is large, individuals in their thirties, the average age within the sample, may have different perspectives, attitudes, and experiences compared to younger or older generations. Additionally, the majority of participants in the study have a technical background, which might affect the generalizability of the findings. Individuals with a technical background may have different expectations and a more critical understanding of AI systems compared to those without technical backgrounds. While valuable insights are gained from studying specific age and educational background characteristics within a particular demographic, it is crucial to recognize potential limitations and carefully consider how applicable the results are to a broader population.

Limitations

It is important to recognize and address the limitations that are associated with this study. The first possible limitation might be given by the majority of participants having a technical

background. Participants with a technical background may possess a different overall perception of AI, potentially having an impact on the study's generalizability. To address this potential bias, controlling for educational background in statistical analyses, a controlled recruiting of participants from diverse educational backgrounds and using a scale that is not self-reported are all factors that might help mitigate this issue. Additionally, there might be a self-selection bias that could play an important role as participants were recruited through online platforms. More specifically, individuals with a particular interest or experience in AI might be overrepresented in the sample as they are the promptest individuals to be interested in this topic. As a result, this bias can have an impact on how applicable and generalizable the findings are since they may not accurately reflect attitudes and reactions to AI mistakes among a wider population.

Moreover, it is important to consider methodological limitations when interpreting the findings. In this study, a non-experimental design was utilized, which involved conducting correlational analysis. This approach restricts the ability to establish direct cause-and-effect relationships between variables. Consequently, it becomes challenging to obtain a comprehensive understanding of how AI literacy, emotional responses, and trust in AI are interconnected.

Finally, the study might present a general trend of possible response bias. This is due the fact that the study relies on self-reported measures for all the variables including AI literacy level, emotional distress to the AI mistakes and trust and attitudes towards AI. For example, participants might have shown social desirability bias by providing responses that are socially acceptable and not the reflection of their true believes. Additionally, participants might have shown response set bias when answering to the questions related to trust and attitude towards AI, after the exposure to AI mistakes. Because they were familiar with the set of questions from the

previous measurement of trust and attitude towards AI before the AI mistakes, they might have given the same responses afterwards, even if the content of the question was different. Finally, situational factors, such as the surrounding environment in which the survey was conducted, might have influenced the responses. To overcome this possible limitation, the integration of qualitative techniques, such as interviews and open-ended questions, might be helpful to capture additional elements, missed in the quantitative assessments.

Recommendations for future research

It is recommended that future research should make efforts to ensure a more diverse representation of participants to enhance the study's external validity and gain a more comprehensive understanding of how individuals from various age groups and educational backgrounds respond to AI mistakes. By encompassing a wider range of demographics, researchers can obtain a deeper insight into the impact of these errors on different populations. In addition, conducting longitudinal studies could provide valuable insights into how people's attitudes and emotional responses towards AI mistakes evolve over time. This dynamic perspective would allow for a better understanding of the long-term effects of AI literacy on individuals' perceptions.

Moreover, it is suggested that researchers implement experimental designs with controlled interventions. These designs would enable them to examine the causal relationships between AI literacy, emotional responses, and trust and attitude towards AI. By manipulating specific variables under controlled conditions, researchers could gain a clearer understanding of how these factors influence one another.

Furthermore, considering the wide range of industries that utilise AI technology, it would be beneficial for future studies to explore context-specific responses to AI mistakes.

Investigating how different sectors may be affected by these errors would provide a more nuanced understanding of their impact on society.

Concludingly, by incorporating these recommendations into future research work, it is possible to examine diverse responses based on demographic differences and gain insights into the evolving attitudes and emotional reactions towards such technological errors over time. Moreover, through experimental designs and context-specific investigations, it could be possible to better understand the complex relationships between AI literacy, emotional responses, and trust in AI across various industries.

Conclusion

This research adds valuable knowledge to the existing understanding of how individuals respond to errors made by AI by exploring the intricate relationships between AI literacy, emotional distress, trust, and attitudes towards AI. Contrary to what was initially hypothesized, a specific level of AI literacy did not influence the emotional distress reported by the participants in response to the AI mistakes. This lack of association may be influenced by other psychological factors like emotional intelligence and coping strategies, as well as contextual factors such as familiarity with situations involving AI. The study also reveals that while lower levels of AI literacy were associated with increased trust in AI, there was a positive correlation between AI literacy and attitudes towards AI. In line with the literature, trust in AI increased after participants were exposed to mistakes made by the AI. This emphasizes how real-life events can impact perceptions of trust and reflects a common tendency among humans to introduce doubt and scepticism when trust is breached. Also, participants' overall attitudes towards AI increased after encountering mistakes, which aligns with individuals' tendency to overly rely on AI systems despite repeated errors possibly due to cognitive biases like

automation bias and sunk-cost fallacy. The study also highlights the importance of considering demographic factors such as age and educational background when interpreting results and acknowledging potential limitations. The findings provide detailed insights into the complex dynamics underlying human responses to mistakes made by artificial intelligence. By putting into practice the recommendations to overcome the possible limitations, future studies might be able to provide a better understanding of which factors can influence human emotional responses to AI mistakes.

References

Alarcon, G. M., Lyons, J. B., Hamdan, I. A., & Jessup, S. A. (2023). Affective responses to trust violations in a Human-Autonomy teaming context: Humans versus Robots. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-023-01017-w

Balakrishnan, J., Dwivedi, Y. K., Hughes, L., & Boy, F. (2021). Enablers and Inhibitors of AI-Powered Voice Assistants: A Dual-Factor approach by integrating the status quo bias and Technology Acceptance model. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10203-y

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think. *Proceedings of the ACM on Human-computer Interaction*, *5*(CSCW1), 1–21. https://doi.org/10.1145/3449287

Carolus, A., Koch, M., Straka, S., Latoschik, M. E., & Wienrich, C. (2023b). MAILS -- Meta AI Literacy Scale: Development and testing of an AI Literacy questionnaire based on Well-Founded Competency Models and Psychological Change- and Meta-Competencies. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2302.09319

Choung, H., David, P., & Ross, A. (2022). Trust in AI and its role in the acceptance of AI technologies. *International Journal of Human-computer Interaction*, *39*(9), 1727–1739. https://doi.org/10.1080/10447318.2022.2050543

De Sá Siqueira, M. A., Müller, B., & Bosse, T. (2023). When Do We Accept Mistakes from Chatbots? The Impact of Human-Like Communication on User Experience in Chatbots That Make Mistakes. *International Journal of Human-Computer Interaction*, 1–11. https://doi.org/10.1080/10447318.2023.2175158

Dikmen, M., & Burns, C. M. (2016). Autonomous Driving in the Real World. *Association for Computing Machinery*. https://doi.org/10.1145/3003715.3005465

Elliott, A. (2019). *The culture of AI: Everyday life and the digital revolution*. Routledge.

Gillath, O., Ai, T., Branicky, Keshmiri, S., Davison, R. B., & Spaulding, R. (2021). Attachment and trust in artificial intelligence. *Computers in Human Behavior*, *115*, 106607. https://doi.org/10.1016/j.chb.2020.106607

Gillespie, N., Lockey, S., Curtis, C., Pool, J., & Akbari, A., 2023, "Trust in Artificial Intelligence: A Global Study", The University of Queensland and KPMG Australia.https://assets.kpmg.com/content/dam/kpmg/au/pdf/2023/trust-in-ai-global-insights-2023.pdf

Gulati, S., Sousa, S., & Lamas, D. (2019). Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology*, *38*(10), 1004–1015. https://doi.org/10.1080/0144929x.2019.1656779

Gulati, S., Sousa, S., & Lamas, D. (2018). Modelling trust in human-like technologies. *Proceedings of the 9th Indian Conference on Human-Computer Interaction*. https://doi.org/10.1145/3297121.3297124

Hidalgo, C. A., Orghian, D., Canals, J. A., De Almeida, F., & Martin, N. (2021). *How humans judge machines*. MIT Press.

Ho, T. M., Mantello, P., & Ho, M. (2023). An analytical framework for studying attitude towards emotional AI: The three-pronged approach. *MethodsX*, *10*, 102149. https://doi.org/10.1016/j.mex.2023.102149

Hristov, K. (2016). Artificial intelligence and the copyright dilemma. *Idea*, *57*, 431.

Kocielnik, R., Amershi, S., & Bennett, P. (2019). Will You Accept an Imperfect AI? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300641

Lee, J. D., & See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

Long, D., & Magerko, B. (2020). What is AI Literacy? Competencies and Design Considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Pp. 1-16)*. https://doi.org/10.1145/3313831.3376727

Lukyanenko, R., Maass, W. & Storey, V.C. Trust in artificial intelligence: From a Foundational Trust Framework to emerging research opportunities. *Electron Markets* **32**, 1993–2020 (2022). https://doi.org/10.1007/s12525-022-00605-4

Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., & Qiao, M. S. (2021). AI Literacy: Definition, Teaching, Evaluation and Ethical Issues. *Proceedings of the Association for Information Science and Technology*, *58*(1), 504–509. https://doi.org/10.1002/pra2.487

Raji, I. D., Kumar, I. E., Horowitz, A., & Selbst, A. D. (2022). The fallacy of AI functionality. *2022 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3531146.3533158

Rosenband, D. L. (2017). Inside Waymo's self-driving car: My favorite transistors. *2017 Symposium on VLSI Circuits*. https://doi.org/10.23919/vlsic.2017.8008500

Rossi, F. (2021). Artificial intelligence: potential benefits and ethical considerations. *Policy Commons*. https://policycommons.net/artifacts/1340375/artificial-intelligence/1950888/

Scantamburlo, T., Cortés, A., Foffano, F., Barrué, C., Distefano, V., Pham, L., & Fabris, A. (2023). Artificial Intelligence across Europe: A Study on Awareness, Attitude and Trust. ArXiv Preprint ArXiv:2308.09979., June 2020, 1–25. http://arxiv.org/abs/2308.09979

Shank, D. B., Graves, C. R., Gott, A., Gamez, P., & Rodriguez, S. (2019). Feeling our way to machine minds: People's emotions when perceiving mind in artificial intelligence. *Computers in Human Behavior*, *98*, 256–266. https://doi.org/10.1016/j.chb.2019.04.001

Schepman, A., & Rodway, P. (2022). The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory Validation and Associations with Personality, Corporate Distrust, and General Trust. *International Journal of Human-Computer Interaction*, *39*(13), 2724–2741. https://doi.org/10.1080/10447318.2022.2085400

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Strauß, S. (2021). Deep Automation Bias: How to tackle a wicked problem of AI? *Big Data and Cognitive Computing*, *5*(2), 18. https://doi.org/10.3390/bdcc5020018

Yampolskiy, R. V. (2016). Taxonomy of pathways to dangerous artificial intelligence. *National Conference on Artificial Intelligence*. https://dblp.uni-trier.de/db/conf/aaai/ethics2016.html#Yampolskiy16

Yampolskiy, R. V., & Spellchecker, M. S. (2016). Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures. *arXiv (Cornell University)*. https://arxiv.org/pdf/1610.07997

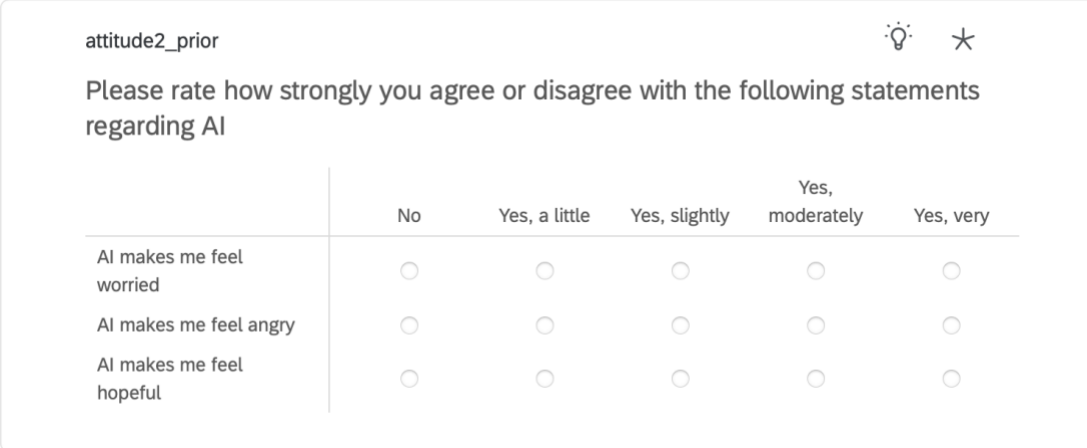
Wagner, D. N. (2021). Economic AI literacy. In *IGI Global eBooks* (pp. 135–152). https://doi.org/10.4018/978-1-7998-5077-9.ch008

Wang, D., Churchill, E. F., Maes, P., Fan, X., Shneiderman, B., Shi, Y., & Wang, Q. (2020). From Human-Human Collaboration to Human-AI Collaboration. *Association for Computing Machinery*. https://doi.org/10.1145/3334480.3381069

Appendices

Appendix A

Attitude towards AI scale


attitude2_prior

Please rate how strongly you agree or disagree with the following statements regarding AI

	No	Yes, a little	Yes, slightly	Yes, moderately	Yes, very
AI makes me feel worried	○	○	○	○	○
AI makes me feel angry	○	○	○	○	○
AI makes me feel hopeful	○	○	○	○	○

Appendix B

AI mistakes scenarios

Plagiarism

"A record label hired an AI songwriter to write lyrics for famous musicians. The AI songwriter has written lyrics for dozens of songs in the past year. However, a journalist later discovers that the AI songwriter has been plagiarizing lyrics from lesser-known artists. Many artists are outraged when they learn about the news."

Inappropriate Content

"A public transport company wanted to create a funny commercial. It decides to commission an advertisement from an AI marketing system that uses a play on the word 'riding'. The resulting

ad, pictured below, causes shock and outrage among members of the public."

Hallucination

"A lawyer at a respected firm used an AI chatbot to find historic cases relevant to his client's lawsuit. The chatbot came up with a list of twelve cases. Later in court it turns out that the chatbots findings were completely made up. Court documents show that half of the submitted cases appear to be bogus judicial decisions with bogus quotes and bogus internal citations."

Harmful content

"An organization that supports people with eating disorders introduced an AI chatbot as a tool that could offer prevention strategies for people with eating disorders, such as anorexia and bulimia. However recently, users started sharing screenshots of their experience with the chatbot via social media. They reported that the bot provided harmful advice. It recommending weight loss, counting calories, and measuring body fat; behaviors that could potentially exacerbate eating disorders. Patients, families, doctors and other experts on eating disorders were left

stunned and bewildered about how a chatbot designed to help people with eating disorders could end up dispensing diet tips instead."

Bias

"To improve their admission process, a university began using a new AI machine-learning system to help make decisions about who gets into its Ph.D. program -- and who doesn't. The algorithm evaluates grades, test scores, and recommendation letters of applicants. An audit revealed that the new algorithm is biased against minority applicants. Critics concerned about diversity, equity and fairness in admissions are angry and say the system exacerbates existing inequality in the field."

Emotional Response

The following questions are about the AI's actions described in the scenario.

	None at all	A little	A moderate amount	A lot	A great deal
How surprising do you find this action?	○	○	○	○	○
How harmful do you find this action?	○	○	○	○	○
How morally wrong do you find this action?	○	○	○	○	○
How emotionally distressing do you find this action?	○	○	○	○	○