Reliability Matters: Exploring the Effect of AI Explanations on Misinformation Detection With a Warning

Haeseung Seo, Sian Lee, Dongwon Lee, Aiping Xiong

The Pennsylvania State University, USA
{hxs378, szl43, dongwon, axx29}@psu.edu

Abstract

To mitigate misinformation on social media, platforms such as Facebook have offered warnings to users based on the detection results of AI systems. With the evolution of AI detection systems, efforts have been devoted to applying explainable AI (XAI) to further increase the transparency of AI decision-making. Nevertheless, few factors have been considered to understand the effectiveness of a *warning with AI explanations* in helping humans detect misinformation. In this study, we report the results of three online human-subject experiments ($N = 2,692$) investigating the framing effect and the impact of an AI system's reliability on the effectiveness of AI warning with explanations. Our findings show that the framing effect is effective for participants' misinformation detection, whereas the AI system's reliability is critical for humans' misinformation detection and participants' trust in the AI system. However, adding the explanations can potentially increase participants' suspicions on miss errors (i.e., false negatives) in the AI system. Furthermore, more trust is shown in the AI warning without explanations condition. We conclude by discussing the implications of our findings.

Introduction

In the context of the COVID-19 pandemic, the overflow of misinformation calls for urgent measures to reduce such misinformation (Bode and Vraga 2021). Many cases have presented how detrimental health-related misinformation is as much as people can sometimes die from the wrong treatment for COVID-19.[1] To mitigate the rapid spread of misinformation on social media, companies such as Meta and Twitter have created warning systems to debunk fake news.[2] Previous studies (Pennycook, Cannon, and Rand 2018; Clayton et al. 2020) have shown that a debunking warning label plays an effective role in mitigating fake news (for a review, see Martel and Rand 2023).

Meanwhile, active efforts have also been devoted to effectively detecting fake news (Shu et al. 2017; Reis et al. 2019; Mosallanezhad et al. 2022). Beyond improving detection models, recent research interest has expanded to explainable artificial intelligence (XAI) to provide an explanation of how an AI system detects fake news to news consumers (Shu et al. 2019) and why a piece of misinformation is false (Dai et al. 2022).

The value of XAI for misinformation mitigation lies in helping users not accept or disseminate it. Empirical studies have been conducted to examine the effectiveness of AI explanations in influencing humans' misinformation detection (Nguyen et al. 2018; Horne et al. 2019; Seo, Xiong, and Lee 2019; Mohseni et al. 2021). For example, Seo et al. (2019) found that AI warning with explanations increased participants' ability to detect fake news when news source was not provided. While Mohseni et al. (2021) found that the extra AI explanations did not reduce participants' belief in misinformation, the explanations helped the participants build appropriate mental models of the AI system.

Despite the promising empirical findings, most of the existing research examines the effectiveness of explanations through options to interact with the AI system, different human behavior (i.e., misinformation detection or sharing), or other factors (e.g., social influence). However, humans' perception and acceptance of an explanation are often shaped by how the problem is framed (Tversky and Kahneman 1981, 1986). Also, prior work showed that AI explanations enhanced participants' misinformation detection but not their trust in the AI system (Seo, Xiong, and Lee 2019). We explore factors that can improve the efficacy of AI warning with explanations in mitigating human's belief in misinformation. We focus on fake COVID-19 news, considering its timely importance. Specifically, we examine the following research questions (RQs).

- **RQ1.** Will a misinformation warning with AI explanations enhance participants' fake news detection compared to the warning only? If so, will a positive framing work better than a negative framing for the explanations?
- **RQ2.** Will the effect of misinformation warning with AI explanations depend on the AI system's reliability?
- **RQ3.** How do the misinformation warning with AI explanations affect participants' trust in the AI system?

We conducted three online experiments by recruiting Amazon Mechanical Turk workers ($N = 2,692$). In Experiment 1, we investigated the effect of explaining how an AI system debunks fake news on participants' detection of

[1] https://www.bbc.com/news/world-53755067

[2] https://www.facebook.com/journalismproject/programs/third-party-fact-checking/new-ratings, https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy

misinformation with a warning (**RQ1**). We proposed credibility explanations in both positive framing (i.e., *POS*) and negative framing (i.e., *NEG*) and examined the framing effect in Experiment 2 (**RQ1**). In Experiment 3, we explored the impact of the AI system's reliability (i.e., whether the AI systems will make a lot of mistakes) (**RQ2**). We also evaluated participants' trust in the AI system (**RQ3**).

The results of our experiments indicate that the AI warning with credibility explanations under the negative framing can reduce humans' perceived accuracy ratings of fake news. Such results underline the necessity to consider the framing effect in designing effective AI explanations for human misinformation detection. However, we find that participants do not always depend upon the warning or the warning with explanations for misinformation detection. They tend to think about miss errors (i.e., false negatives) of the AI system. Moreover, the system's reliability is critical to address such suspicion. Those results highlight the importance of informing users of the AI system's reliability and possible error types. Finally, although participants' misinformation detection can be influenced by the AI explanations, they show more trust in the warning itself. Moreover, such trust does not depend on the AI system's reliability. Our main contributions are summarized as follows.

- We empirically examine the framing effect on misinformation warning with explanations upon humans' misinformation detection.
- We present evidence showing the effectiveness of high system reliability in humans' misinformation detection and their trust in AI systems.
- We provide implications for researchers and practitioners in designing AI explanations to mitigate misinformation on social media platforms.

Related Work

Misinformation and Its Correction

Misinformation and fake news have been broadly used as umbrella terms to refer to false or fabricated information written and published online (Lazer et al. 2018; Wu et al. 2019). In this work, we use these two terms interchangeably.

The rampant spread of misinformation on social media has attracted researchers (Bode and Vraga 2018; Clayton et al. 2020; Pennycook, Cannon, and Rand 2018; Seo, Xiong, and Lee 2019) and social media platforms (e.g., Facebook) to explore effective ways to debunk misinformation. A widely adopted debunking approach is to apply warning tags, labels, or indicators, during the misinformation presentation after fact-checking by professional organizations or artificial intelligence (AI). Empirical user studies reveal that those warnings are generally effective in reducing participants' belief in misinformation (Clayton et al. 2020; Yaqub et al. 2020; Jia et al. 2022; Kreps and Kriner 2022; Lu et al. 2022). Yet, the efficacy of the warnings can be impacted by factors such as warning specificity (e.g., general warnings introduce bias, reducing belief in real news), warning design (e.g., simple and precise warning language), the source of the warnings (e.g., fact checker and community), and extra fact-checking details.

Explainable Artificial Intelligence (XAI) and Misinformation Correction

Recently, misinformation detection work has been proposed by leveraging machine learning (ML) or AI algorithms (Shu et al. 2017; Cui et al. 2019; Shu et al. 2019; Zhou and Zafarani 2020). Given the challenges of automated fact checking and possible biases introduced in ML dataset and training (Guo, Schlichtkrull, and Vlachos 2022), it is essential to enhance the transparency of ML/AI-based misinformation detection beyond the warning indicators. In line with this idea, XAI aims to make AI system results understandable or interpretable to humans using explanations (Adadi and Berrada 2018; Gunning and Aha 2019; Arrieta et al. 2020).

A few studies have been conducted to understand AI-based misinformation warning with explanations (Seo, Xiong, and Lee 2019; Horne et al. 2019; Mohseni et al. 2021; Epstein et al. 2022). However, those studies either did not explain concretely how an AI system derived each prediction (i.e., lack of specificity at the local level) or how the AI system behaves in general (i.e., lack of system performance at the global level). Results of those studies also revealed mixed findings. While the effect of ML/AI warning with explanations was evident in some studies (Seo, Xiong, and Lee 2019; Epstein et al. 2022), it was not obtained in others (Mohseni et al. 2021). Those initial efforts have provided a preliminary understanding of the effect of AI misinformation warning with explanations. To the best of our knowledge, how to design AI warning with explanations for effective misinformation mitigation has not been well understood. Our study aims to fill the gap by examining the framing effect (Tversky and Kahneman 1981) and the impact of AI system's reliability (Lee and See 2004).

Framing Effects

In XAI, a bar chart has been often used to explain features that impact AI decision-making (Cheng et al. 2019; Wang and Yin 2021). Seo et al. (2019) also investigated its use to explain misinformation detection results of an AI system. However, the effect of explanations was only evident when compared to a control without warning or explanation. We complement and extend the use of bar charts by exploring the framing effect. Tversky and Kahneman addressed the framing effect first by explaining that people's willingness to take risks can depend on how options are presented (Tversky and Kahneman 1981, 1986). In the usable privacy literature, Choe et al. (2013) investigated the visual framing effect in mobile app's privacy information. Their results suggested the effect of a positively-framed privacy rating icon in nudging people away from privacy-invasive apps. Greene and Murphy (2021) investigated the framing effect of a general misinformation warning message and obtained no impact of the framing effect, which could be due to the ineffectiveness of general misinformation warning (Clayton et al. 2020). Despite different domains and varied framing designs, these studies throw insights that can be applied to our study. From the framing effect point of view, we propose to compare a negative framing and a positive framing of bar chart in addition to the warning message against fake news.

Credibility for Misinformation Correction

Even though it is hard to find a universal agreement on the concept of credibility across different fields (Savolainen 2021), credibility has been used as a primary criterion to measure the quality of information on the web (Flanagin and Metzger 2000) as well as traditional mass media (Gaziano and McGrath 1986). Credibility can be understood in terms of believability, trust, reliability, accuracy, fairness, and objectivity (Savolainen 2021); therefore, it is naturally emphasized in human misinformation detection. A number of studies have investigated the credibility of information on social media (Westerman, Spence, and Van Der Heide 2014; Lin, Spence, and Lachlan 2016; Savolainen 2021). Among them, Savolainen (2021) suggested a conceptual framework of credibility by dividing it into two approaches: 1) the credibility of the author (i.e., source) and 2) the credibility of misinformation content. Molina et al. (2021) also identified extra features of fake news such as user comments, which are interpretable and useful for humans to evaluate news veracity. Based on these studies, we select three factors (i.e., source, content, and user comments) and choose credibility as the gauge to quantify the veracity of misinformation in the proposed bar charts. We use falsity (Seo et al. 2021) in the negative framing.

Trust in AI Systems

The concept of trust has been defined differently across various fields (Mayer, Davis, and Schoorman 1995; Hoff and Bashir 2015; Chancey et al. 2017). For example, Mayer et al. define trust as a willingness to accept vulnerability. Trust in information systems indicates self-assurance by assessment of risks and alternatives (Pieters 2011). Furthermore, trust has been considered as a behavior that can have an impact on users' dependence on automation (Lee and See 2004). Machine learning researchers have also paid attention to the importance of trust linked to model justification issues (Ribeiro, Singh, and Guestrin 2016; Lipton 2018; Toreini et al. 2020).

An explanation is expected to increase trust through contributing to enhanced transparency of AI systems (Lipton 2018). However, building human trust in algorithmic systems can be a challenging task (Lee 2018). Seo et al. (2019) conducted user studies investigating the effects of AI warning with explanations. While the AI warning with specific explanations using bar charts enhanced participants' fake news detection, it did not increase their trust in the AI system. Using AI warning with a general explanation describing how (i.e., the process) an AI warning label was created, Epstein et al. (2022) replicated Seo et al.'s findings when examining participants' misinformation sharing.

Reliability of AI Systems. Reliability is regarded as one of the bases of trust (Lee and See 2004). Previous studies showed that users' trust in a warning system can be impacted by the reliability of the system (Chen et al. 2018). Several researchers demonstrated the impact of reliability information in building users' trust in automated systems (Dzindolet et al. 2003; Chancey et al. 2017). Dzindolet et al. (2003) confirmed that trust in automation can increase with information about why a decision of an automated system might err. In Chancey et al.'s study, the high-reliability system got more trust from the participants than the low-reliability system. Furthermore, Kocielnik et al. (2019) explored the impacts of different types of errors an AI system made. The results showed that how users reacted largely depended on how the AI system behaved in general. Based on these findings, we investigate the impact of an AI system's reliability on the efficacy of AI warning with explanations by comparing a high-reliability system with a low-reliability system. To the best of our knowledge, our study is the first study to examine the reliability information of an AI system in terms of explanation-based-warnings.

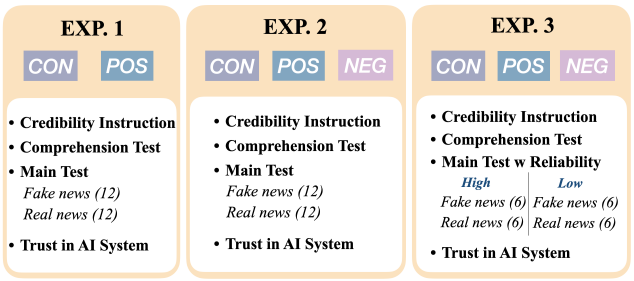


Figure 1: An overview of the experiment design. Experiments 1 and 2 investigate the framing effect. Experiment 3 focuses on reliability. *CON* means control (the warning-only condition). *POS* and *NEG* mean the warning with positively- and negatively-framed explanations conditions, respectively.

The Present Study

We conducted three online experiments to investigate the effect of explaining how an AI system debunks fake news can enhance the effectiveness of fake news warnings in the context of social media platforms. As shown in Figure 1, we investigated the effect of a warning with credibility explanations compared with the warning-only condition in Experiment 1 (**RQ1**). In Experiment 2, we explored whether the framing effect matters for a warning with explanations by comparing positively-framed explanations (i.e., credibility) and negatively-framed explanations (i.e., falsity) (**RQ1**). In Experiment 3, we examined the effect of AI system's reliability on the proposed explanations (**RQ2**). In each experiment, participants evaluated 24 pieces of news (half fake) by answering their perceived accuracy rating and confidence in the perceived accuracy decisions. We also measured their trust in the AI systems across all experiments (**RQ3**).

Participants

We recruited participants on Amazon Mechanical Turk (MTurk) through the Human Intelligent Task (HIT) for all experiments. The HITs included the task description, and workers were able to decide whether they would like to perform the task. In each experiment, we required the workers to be those who (1) are at least 18 years old, (2) live in the U.S., and (3) have finished more than 100 HITs with a HIT approval rate of at least 95%. Across experiments, MTurk workers were only allowed to participate in our study once.

Items	Options	EXP.1 (N=710)	EXP.2 (N=1014)	EXP.3 (N=968)
Gender	Male	51.8%	42.2%	38.7%
	Female	47.6%	57.4%	60.1%
	PNA	0.6%	0.4%	1.1%
Age	18-29	21.3%	24.3%	20.0%
	30-39	34.2%	33.0%	33.6%
	40-49	25.1%	24.5%	24.8%
	50-59	14.4%	12.9%	13.5%
	60-79	5.1%	5.2%	8.0%
Race	Caucasian	78.9%	79.6%	75.2%
	African American	10.7%	9.0%	11.1%
	Hispanic	3.7%	4.8%	4.8%
	Asian	3.5%	4.0%	5.7%
	Other	2.9%	1.9%	1.5%
	PNA	0.3%	0.7%	0.7%
Education	High school	4.9%	7.4%	8.9%
	College credit	12.3%	16.4%	27.2%
	Bachelor	58.7%	53.0%	38.5%
	Master	20.8%	19.2%	17.1%
	Doctor	1.8%	1.7%	3.7%
	Other	1.1%	2.1%	3.9%
	PNA	0.3%	0.3%	0.6%
AI/ML experience	Not at all	16.1%	27.0%	46.3%
	Novice	19.2%	25.8%	37.1%
	Intermediate	28.0%	21.3%	12.6%
	Competent	27.3%	18.0%	3.8%
	Expert	9.4%	7.8%	0.1%

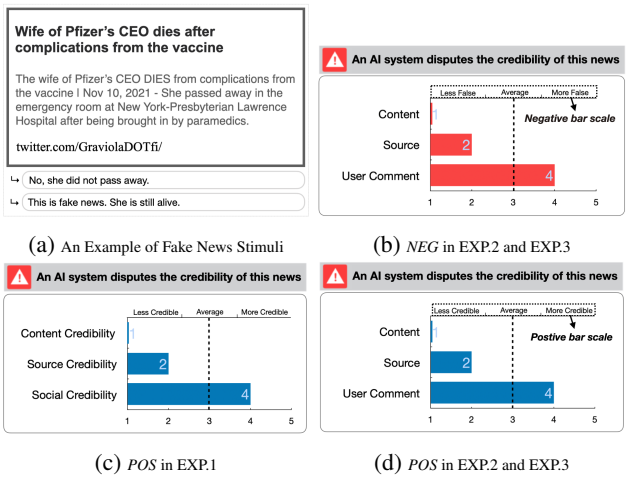
Table 1: Demographic information of the participants in the three experiments. EXP denotes experiment. PNA refers to Prefer Not to Answer.

We recruited 1,246 (EXP.1), 1,686 (EXP.2), and 1,196 (EXP.3) participants. We manually checked the responses and ensured that there was no duplicate participation across experiments. We also removed respondents (1) outside of the U.S., (2) with duplicate IPs, (3) failed the comprehension test, (4) failed the attention check, and (5) with completion time shorter than 3 min (average median completion time: 15 min). The number of participants we accepted was 710 (EXP.1), 1014 (EXP.2), and 968 (EXP.3), respectively. The high exclusion rate in Experiments 1 and 2 was to ensure our data quality,[3] which was necessary given the concerns on the MTurk platform (Peer et al. 2022). In Experiment 3, we used the MTurk toolkit CloudResearch provides to automatically exclude low-quality workers (Litman, Robinson, and Abberbock 2017; Hauser et al. 2022). Based on an hourly rate of $7.5, participants were paid $1.8 for completing a study. Participants' demographic information is shown in Table 1.

Materials

News Stimuli. We selected 25 news articles about COVID-19 released from September to November 2021. Twelve pieces of fake news were searched from *snopes.com* or *politifact.com*, both of which are representative fact-checking websites. Thirteen pieces of real news were selected from major news platforms such as *cnn.com* or *apnews.com*. A piece of real news was for an attention check (Hauser and Schwarz 2016).

[3]Participants who answered more than half of the questions correctly passed the comprehension test. See Exclusion Details in the supplementary materials: http://tiny.cc/seoetal24supp



(a) An Example of Fake News Stimuli

(b) *NEG* in EXP.2 and EXP.3

(c) *POS* in EXP.1

(d) *POS* in EXP.2 and EXP.3

Figure 2: A piece of fake news headline, including the news title, a snippet of the news article, and the source, is presented concurrently with two user comments *(a)*. For each fake headline, a warning label is shown below the user comments. Each bar chart is presented below the warning label as positively- (*POS*) or negatively-framed (*NEG*) explanations (*(b)-(d)*). Each bar name is shortened to avoid redundancy/conflict with bar scales in Experiments (EXP) 2 & 3.

In all experiments, we present a piece of real or fake news in the form of a news headline with two fictional users' comments (see Figure 2(*a*)). Each news headline is composed of a title, a snippet of the article, and a source. For the source, real news headlines have URLs from major news platforms where the real news was excerpted. Fake news headlines have social media URLs where the misinformation was posted. The users' comments for fake news have negation-style sentences debunking the misinformation, and the comments for real news have a neutral tone, not directly pointing out information veracity (Seo et al. 2021).

Warning and Explanation Interfaces. In our design, we assume that each piece of fake news has been detected by an AI system. Thus, a warning label is shown for each piece of fake news, describing that the fake news has been disputed by an AI system (see Figure 2). There is a baseline condition in each experiment, in which we present the warning label only. Considering the robust effect of warning labels (Clayton et al. 2020) and our primary interest in the effect of AI explanations, we omit a condition without warning and define the baseline with a warning label as the control (*CON*).

Seo, Xiong, and Lee (2019) proposed a bar chart presenting factors that AI systems consider when debunking fake news. The authors presented the bar chart as an explanation for AI decision-making and found impacts of the explanation on participants' fake news evaluation. We adapt their design and create two types of explanations: positive framing (*POS*) and negative framing (*NEG*). Figures 2(*b*) and 2(*d*) show an example for each type, illustrating the fake news credibility (*POS*) and falsity (*NEG*), respectively.

Positive Framing (*POS*). We present three factors that our hypothetical AI system considers, including *content* credibility based on the news title and contents, *source* credibility based on the news source, and *social* credibility based on users' comments. A filled blue bar graph accompanies each factor, and the length of each bar indicates the credibility score that the AI system derived for evaluating the factor. For the bar graphs, "More Credible," "Average," and "Less Credible" are marked on the top of the bar graph panel, and numbers "1" through "5" are marked on the bottom of the panel (see Figure 2). There is an outline frame that indicates the possible maximum score of "5." For a score of "1", a short blue bar is displayed. For a score of "5", the blue bar is extended toward the rightmost of the bar graph. Thus, the range of the score is clear, and a direct visual comparison is enabled among the bars (Cleveland 1985).

We create 12 bar charts to be added to the 12 pieces of fake news. We score each factor using a 5-point score. We use either 1, 2, 4 or 1, 2, 5 for value combination avoiding 3, a neutral number. Each factor shows its high credibility (i.e., 4 or 5) four times among the 12 pieces of fake news. Moreover, we separate the 12 pieces of fake news into three sets and implement a Latin-square design to counterbalance the credibility value combinations across different sets.[4]

Negative Framing (*NEG*). In addition to the positive framing using blue bars, we propose explanations using a negative framing (Choe et al. 2013). The interface is the same as the credibility explanations, except that we change the wordings of the bar scale (e.g., "Less Credible" to "Less False") and the color of bar graphs from blue to red (see Figure 2). Instead of an equivalent framing, we apply the same score set to the falsity explanations. Like the credibility scores, each factor has the highest value four times for the falsity scores. Thus, compared to the credibility explanations using positive framing, fake news with the falsity explanations using negative framing has a lower falsity score (see *panels (b)* and *(d)* in Figure 2). We also remove "Credibility" in the factor description of the *POS* condition in Experiment 2 to reduce redundancy with the bar scale and make the interface comparable to that of the *NEG* condition.

Procedure

Qualtrics was used to design our online studies. After informed consent, participants were randomly assigned to one condition in each experiment. We first described our hypothetical AI system and asked participants questions to check their comprehension of our design. We asked two common questions for all conditions but added three extra questions for the *POS* and *NEG* conditions to check participants' comprehension of each bar chart category. Then, all news stimuli were presented in a randomized order. Twelve of them included fake news, and 13 of them included real news. One of the real news was for an attention check. We provided participants with specific instructions on how to answer the attention-check question (Hauser and Schwarz 2016). For any participants who failed to follow the instructions, their survey was terminated immediately. We paid those participants a base payment of $0.5.

We asked two questions investigating participants' acceptance of the "claim" in each news stimulus. First, participants rated their perceived accuracy rating of the news claim, "How accurate is the claim in the above news?" using a 7-point scale (1: very inaccurate, 7: very accurate). Then they answered the other question about their confidence in their perceived accuracy rating, "How confident are you in answering the question above?" using another 7-point scale (1: not confident at all, 7: fully confident).

After answering questions for the 25 pieces of news, there was a post-session questionnaire. We asked participants four questions to measure their trust in the AI system disputing the fake news, including "I trust the AI System when making judgments about news veracity," " The AI warning is informative when I make judgments about news veracity," "The AI warning is helpful when I make judgments about news veracity," and "I would like to see the AI system implemented on social media." Participants rated their agreement with each question using a 7-point Likert scale, with "1" indicating "strongly disagree" and "7" indicating "strongly agree." In the end, participants filled in their demographic information, including age, gender, ethnicity, and education, and their experience in AI or machine learning.

Results

The three critical dependent measures we assessed were participants' *perceived accuracy rating*, *confidence in accuracy rating*, and *trust in the AI system*. Descriptive statistics for the measures are shown in Figure 3. Across three experiments, we analyzed responses from 710, 1014, and 968 participants using SPSS version 29.

Experiment 1

The goal of Experiment 1 was to verify whether our proposed warning with explanations design can enhance participants' misinformation detection compared to the warning-only condition (**RQ1**). We manipulated two factors with condition between subjects and news veracity within subjects. To quantify the effects, we entered perceived accuracy and confidence results into 2 (*news veracity*: real, fake) $\times$ 2 (*condition*: *CON*, *POS*) mixed analysis of variance (ANOVA) tests with α = .05. We chose ANOVA since it is robust against violations of the underlying assumption of normally distributed data (Norman 2010). Post-hoc tests with Bonferroni correction were performed, testing all pairwise comparisons with corrected p values for possible inflation. The participant counts included for data analysis are 413 in the *CON* condition and 297 in the *POS* condition.

Perceived Accuracy Rating. As shown in Figure 3, participants clearly distinguished the real news (5.49) from the fake news (3.65), $F_{(1,708)} = 560.03$, $p < .001$, $\eta_p^2 = .442$. Yet, the two-way interaction of *news veracity* $\times$ *condition* was not significant, $F < 1.0$. Instead, participants in the *POS* condition (4.47) gave lower rating than those in

[4]We focus on the credibility/falsity value combinations but do not control the value alignment for each factor. Our post-hoc analyses on the perceived accuracy rating of fake news across the three factors show no significant differences, suggesting limited impacts.

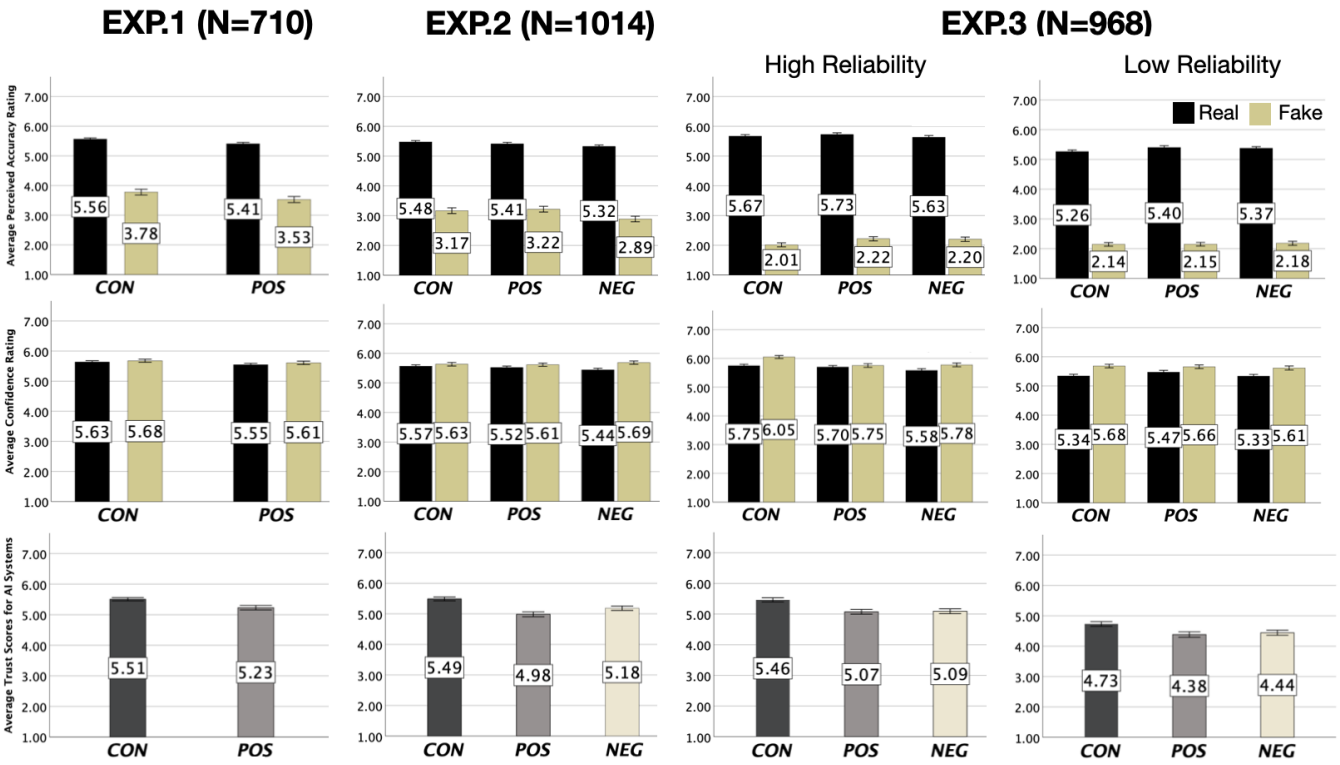


Figure 3: Mean values of perceived accuracy ratings (top row), confidence in the perceived accuracy ratings (central row), and trust in the AI system (bottom row) across each condition in the three experiments (1st column: Experiment (EXP.) 1, 2nd column: EXP.2, 3rd and 4th columns: EXP.3). *CON*: control condition showing only the warning label, *POS* and *NEG*: conditions showing warning with explanations in a *positive* and *negative* framing. Error bars represent ± one standard error.

the *CON* condition (4.67) regardless of the news veracity, $F_{(1,708)} = 6.74$, $p = .010$, $\eta_p^2 = .009$. Post-hoc tests revealed the main effect of *condition* for the real news, $F_{(1,708)} = 6.35$, $p = .012$, $\eta_p^2 = .009$, but not the fake news, $F_{(1,708)} = 3.07$, $p = .080$, $\eta_p^2 = .004$. Thus, contrary to expectation, the explanations in the *POS* condition showed less impact on participants' fake news evaluation than their real news evaluation.

Confidence in Accuracy Rating. Participants were confident in their perceived accuracy ratings in general (see Figure 3). They gave similar confidence ratings in their decisions of the fake news (5.64) and the real news (5.59), $F_{(1,708)} = 2.45$, $p = .118$, $\eta_p^2 = .003$. They also showed comparable confidence in both conditions (*CON*: 5.66 and *POS*: 5.58), $F_{(1,708)} = 1.52$, $p = .219$, $\eta_p^2 = .002$. The two-way interaction of *news veracity* × *condition* was not significant either, $F < 1.0$. Thus, the extra explanations did not impact participants' confidence in fake news detection.

Trust in the AI System. We calculated the average ratings of the four questions asking about participants' trust in the AI system. Participants in the *CON* condition showed higher trust in the AI system (5.51) than those in the *POS* condition (5.23), $F_{(1,708)} = 9.67$, $p = .002$, $\eta_p^2 = .013$.

Influential Credibility. To understand the relative weighting of the three factors in the explanations, we analyzed the accuracy rating of each factor with high credibility using ANOVA. Mean values of content credibility (3.59), source credibility (3.51), and social credibility (3.48) showed no statistical difference, $F_{(2,296)} = 2.82$, $p = .094$, $\eta_p^2 = .009$.

Summary and Discussion. In Experiment 1, we observed that the participants who were exposed to the extra explanations in the *POS* condition decreased their perceived accuracy ratings in general, particularly on the real news, which was in the opposite direction to the expected effect.

We gauged the credibility of each factor positively ("more" means "better") and presented the credibility score using a blue color. However, the AI system in our study is mainly debunking fake news. According to the framing effect, human decision-making in risky contexts is influenced by how a problem is framed (Tversky and Kahneman 1981). Considerations of compatibility indicate that positive dimensions are weighted more when the task is to accept, whereas negative dimensions are weighted more when the task is to reject (Shafir 1993). Thus, we conjecture that the ineffectiveness of the extra explanations on enhancing fake news detection could be due to the incompatibility between its positive framing and the debunking of misinformation.

The factors listed in the explanations are interpretable and applicable to human decision-making. The reduced accuracy rating in the *POS* condition indicates that participants might have considered those factors for the real news evalu-

ation. Also, we included two pieces of real news from *CNN*, which has been viewed as a source of fake news (Mastrine 2024). Thus, the extra explanation might have "helped" participants detect "fake news missed" by the AI system, and consequently reduced their trust in the system.

Experiment 2

The purpose of Experiment 2 was to investigate the framing effect by comparing positively-framed explanations with negatively-framed explanations (**RQ1**). The experimental setting was similar to Experiment 1, except that we added *NEG* as another between-subjects condition. The *POS* and *NEG* conditions were the same except for the wording of bar gauge and the color of the bar chart (see Figures 1 and 2 for the details). All data analyses were conducted in the same way as Experiment 1, except that perceived accuracy rating and confidence rating were entered into 2 (*news veracity*: real, fake) × 3 (*condition*: *CON*, *POS*, *NEG*) ANOVAs. The number of participants included for data analyses is as follows: 390 (*CON*), 309 (*POS*), and 315 (*NEG*).

Perceived Accuracy Rating. The average results of the real and fake news for each condition are shown in Figure 3. Same as Experiment 1, there were main effects of *news veracity*, $F_{(1,1011)} = 1353.97$, $p < .001$, $\eta_p^2 = .573$, and *condition*, $F_{(2,1011)} = 5.25$, $p = .005$, $\eta_p^2 = .010$, but not their two-way interaction, $F_{(2,1011)} = 1.15$, $p = .316$, $\eta_p^2 = .002$. Specifically, participants gave higher accuracy ratings for the real news (5.40) than for the fake news (3.09). The average rating of the *NEG* condition (4.11) was smaller than those of the *CON* condition (4.32, $p = .010$) and the *POS* condition (4.31, $p = .022$), respectively. Post-hoc tests revealed that the main effect of condition was only significant for the fake news (*CON*: 3.17, *POS*: 3.22, *NEG*: 2.89), $F_{(2,1011)} = 3.24$, $p = .039$, $\eta_p^2 = .006$, but not the real news (*CON*: 5.48, *POS*: 5.41, *NEG*: 5.33), $F_{(2,1011)} = 2.73$, $p = .066$, $\eta_p^2 = .005$. Thus, participants in the *NEG* condition only showed a non-significant trend of reducing their perceived accuracy for the real news. Consequently, the main effect of condition across the two veracity levels showed an opposite pattern to what we obtained in the *POS* condition of Experiment 1, indicating the framing effect.

Confidence in Accuracy Rating. Figure 3 presents the average confidence rating of each condition. Different from Experiment 1, participants were more confident in their ratings of fake news (5.64) than real news (5.51), $F_{(1,1011)} = 21.70$, $p < .001$, $\eta_p^2 = .021$. Although their confidence ratings were similar across conditions (*CON*: 5.60; *POS*: 5.57; *NEG*: 5.56), $F < 1.0$, the interaction of *news veracity* × *conditions* was significant, $F_{(2,1011)} = 3.66$, $p = .026$, $\eta_p^2 = .007$. Post-hoc tests on the main effect of the condition were not significant in either veracity, $F_s < 1.77$. Thus, as shown by the numerically highest rating of the fake news but the numerically lowest rating of the real news in Figure 3, the interaction mainly suggests the main effect of news veracity for the *NEG* condition but not the other conditions.

Trust in the AI System. Participants' trust scores varied across conditions, $F_{(2,1011)} = 13.54$, $p < .001$, $\eta_p^2 = .026$. Same as Experiment 1, participants in the *CON* condition gave the highest trust score (5.49), which was significantly higher than those in the *NEG* condition (5.18, $p = .006$) and the *POS* condition (4.98, $p < .001$). The trust scores between the latter two conditions were similar ($p = .163$).

Influential Credibility. We analyzed the accuracy rating of each factor with high credibility/falsity using ANOVA with 3 (*factor*: content, source, user comments) as a within-subjects variable and 2 (*condition*: *POS*, *NEG*) as a between-subjects variable. The main effects of *factor*, $F_{(2,622)} = 4.25$, $p = .040$, $\eta_p^2 = .007$, *condition*, $F_{(1,622)} = 6.02$, $p = .014$, $\eta_p^2 = .010$, and their interaction, $F_{(2,622)} = 26.76$, $p < .001$, $\eta_p^2 = .041$, were all significant. Post-hoc pairwise comparisons revealed that the factor of user comments (3.12) showed a higher accuracy rating than the factor of content (3.01, $p = .042$), but not significantly different with the factor of source (3.03, $p = .067$). Also, the latter two showed no significant difference ($p > .999$). Furthermore, participants in the *NEG* condition (2.89) gave lower accuracy ratings than those in the *POS* condition (3.22), indicating the effect of negative framing. Moreover, such an effect was *factor-dependent*. Post-hoc pairwise comparisons revealed the framing effect on content (*POS*: 3.31, *NEG*: 2.72, $p < .001$) and source (*POS*: 3.21, *NEG*: 2.84, $p = .009$), but not user comments (*POS*: 3.13, *NEG*: 3.10, $p = .810$). The user comments directly debunk misinformation without specifying who the commenters are, which might have resulted in the nonsignificant effect.

Summary and Discussion. In Experiment 2, we further examined the effects of warning with explanations using negative framing (*NEG*) and positive framing (*POS*). We found that the negative framing was effective in reducing participants' perceived accuracy of fake news. Such results highlight that it is essential to consider the framing effect in designing explanation interfaces for the debunking of misinformation. As suggested by the non-significantly reduced perceived accuracy for real news, participants in the *NEG* condition seemed to have not relied on the AI system for the evaluation. Same as Experiment 1, they might have leveraged factors learned from the explanations and made their own decisions. Such dependence gap between real and fake news evaluations was further implied by the numerically highest confidence rating of fake news but the numerically lowest confidence rating of real news in the *NEG* condition.

Experiment 3

Across Experiments 1 and 2, we observed that the participants did not always depend upon the warning or the warning with explanations for their accuracy ratings. Especially, the participants were suspicious when the AI system "failed" to tag a piece of "fake" news (i.e., a system error of *miss*). Such results suggest that beyond *local* explanations for each specific decision, participants have concerns about the AI system's performance at a *global* level. Thus, in Experiment 3, we also described the *reliability* of the AI system's fake news detection. We varied it on two levels (low vs. high) and examined its impacts on the three dependent measures.

The experimental design was the same as Experiment 2 except as noted. We varied the reliability description within subjects but counterbalanced the order of the two reliabilities between subjects. At the beginning of each phase, the reliability information of the AI system was presented. We adapted the instructions of (Chancey et al. 2017). In the low-reliability phase, we presented, "In this phase, the AI system to detect fake news could be pretty unreliable, so it probably will make a lot of mistakes." In the high-reliability phase, "In this phase, the AI system to detect fake news would be pretty reliable, so it probably will NOT make a lot of mistakes." was shown. We used the same news stimuli as Experiments 1 and 2 but split it into two sets (see Figure 1). In each reliability phase, six pieces of real and six pieces of fake news were presented, respectively. The two sets were chosen to have a similar distribution based on the perceived accuracy ratings of each piece of news in Experiment 2. The number of participants included for data analyses is as follows: 359 (*CON*), 300 (*POS*), and 309 (*NEG*).

Perceived Accuracy Rating. Results of the average perceived accuracy ratings are shown in Figure 3. We ran mixed ANOVAs with 3 (*condition*: *CON*, *POS*, *NEG*) $\times$ 2 (*news veracity*: real, fake) $\times$ 2 (*reliability*: low, high). Same as the prior experiments, participants clearly distinguished the real news (5.51) from the fake news (2.15), $F_{(1,965)} = 4021.36$, $p < .001$, $\eta_p^2 = .806$. Yet, the perceived accuracy rating of fake news in Experiment 3 was much lower than those in the prior experiments, which might be due to the time gap between experiments. We will discuss it in later sections.

The main effect of *reliability*, $F_{(1,965)} = 56.92, p < .001$, $\eta_p^2 = .056$, and the interaction of *news veracity* $\times$ *reliability*, $F_{(1,965)} = 67.57$, $p < .001$, $\eta_p^2 = .065$, were significant. Compared to the low-reliability condition, participants in the high-reliability condition increased their accuracy ratings for the real news (low: 5.35 vs. high: 5.68, $p < .001$), but their accuracy ratings for fake news showed no significant differences (low: 2.16 vs high: 2.14, $p = .656$).

The three-way interaction of *news veracity* $\times$ *reliability* $\times$ *condition* was also significant, $F_{(2,965)} = 5.72, p = .003$, $\eta_p^2 = .012$. For the real news, the increased accuracy ratings in the high-reliability conditions were similar across conditions ($p_s < .001$). However, for the fake news, only the participants in the *CON* condition gave lower perceived accuracy ratings when the AI system's reliability became higher ($p = .008$) but not those in the other two conditions ($p_s > .203$). Thus, when the AI system became more reliable, it addressed the participants' concerns about possible miss errors in the two explanation conditions. Moreover, regardless of the news veracity, it enhanced the participants' accuracy judgment in the warning-only condition in general. All the other effects were not significant, $F_s < 2.94$.

Confidence in Accuracy Rating. Same as Experiment 2, the main effect of *news veracity*, $F_{(1,965)} = 71.66$, $p < .001$, $\eta_p^2 = .069$, and the interaction of *news veracity* $\times$ *condition*, $F_{(2,965)} = 4.91, p = .008$, $\eta_p^2 = .010$, were significant. The participants showed higher confidence in their ratings of the fake news (5.76) than the real news (5.53). Such a gap was more evident in the *CON* ($p < .001$) and *NEG* ($p < .001$) conditions than in the *POS* ($p = .013$) condition.

Participants were more confident in their accuracy ratings when the AI system's reliability became higher, $F_{(1,965)} = 135.65$, $p < .001$, $\eta_p^2 = .123$. There was also an interactions of *reliability* $\times$ *veracity*, $F_{(1,965)} = 7.21$, $p = .007$, $\eta_p^2 = .007$, qualifying the impact of reliability on the confidence rating gap between real and fake news. As shown in Figure 3, the gap was larger when the system reliability was low (0.27) than when it was high (0.18). The interactions of *reliability* $\times$ *condition* was also significant, $F_{(2,965)} = 10.87$, $p < .001$, $\eta_p^2 = .022$. When the AI system's reliability was low, participants' confidence ratings were similar across the conditions ($p_s > .796$). When the AI system became more reliable, participants' confidence ratings varied across the conditions. Specifically, participants' rating in the *CON* condition (5.90) was higher than that in the *NEG* condition (5.68, $p = .007$), both of which showed no significant difference compared to that of the *POS* condition (5.73, $p_s > .052$). Such results are in agreement with the better accuracy judgment obtained for the *CON* condition. No other significant effects were observed, $F_s < 1.67$.

Trust in the AI System. Participants' trust scores varied across conditions, $F_{(2,965)} = 8.41$, $p < .001$, $\eta_p^2 = .017$. Post-hoc pairwise comparisons showed that the participants trust the *CON* condition the most (5.10), followed by the *NEG* condition (4.77, $p = .003$) and the *POS* condition (4.73, $p < .001$). Participants gave a higher trust score for the system of high reliability (5.21) than that of low reliability (4.52), $F_{(1,965)} = 331.51$, $p < .001$, $\eta_p^2 = .256$. There was no interaction of *reliability* $\times$ *condition*, $F < 1.0$.

Influential Credibility. We analyzed the perceived accuracy rating results using the same method as Experiment 2. Only the main effect of *factor* was significant, $F_{(2,607)} = 9.98$, $p < .001$, $\eta_p^2 = .016$. Post-hoc pairwise comparisons revealed that the factor of *user comments* (2.10) showed lower accuracy rating ($p_s < .018$) than the factors of *source* (2.20) and *content* (2.27), but the latter two showed no significant difference ($p = .221$). Thus, when the AI system reliability was varied, participants might have relied more on the direct debunking in user comments for the evaluation.

Summary and Discussion. The results of Experiment 3 verified that the AI system's reliability is critical to addressing participants' suspicions about the AI system's decision on real news in the two explanation conditions. Moreover, we observed that impacts of the system reliability on the accuracy rating of fake news varied across the conditions: participants in the *CON* condition became less worried about mistakenly labeled fake claims (i.e., a system error of *false alarm*) when the AI system's reliability increased, whereas no significant difference was obtained in the two explanations conditions, implying the influence of other factors. While confidence results did not yield the same pattern as the accuracy rating results, they generally agreed with each other. For the trust measure, we obtained the main effects of condition and reliability but not their interaction, suggesting two different bases for participants' trust in the AI system.

General Discussion

Across three experiments, we evaluated the effect of explaining how an AI system debunks fake news on humans' detection of misinformation with a warning. We proposed credibility explanations in both positive framing (i.e., *POS*) and negative framing (i.e., *NEG*) and examined the framing effect in Experiment 2. In Experiment 3, we further varied the AI system's reliability (i.e., descriptions about whether or not the AI system will make a lot of mistakes).

We obtained evidence of the framing effect: participants who were exposed to the credibility explanation under the negative framing gave lower accuracy ratings for fake news and tended to be more confident in their accuracy decisions than those exposed to the explanation under the positive framing (**RQ1**). However, the participants in the two explanation conditions did not always depend on the warning or warning with explanations for detecting misinformation. In particular, they became suspicious about false negatives (i.e., the AI system error due to *miss*). Furthermore, we found that the system's reliability was critical to address such suspicions of the participants (**RQ2**). Across the three experiments, we obtained two bases of participants' trust in the AI system (i.e., more trust in the warning and more trust when the AI system became more reliable, **RQ3**).

Positive Versus Negative Frames When Explaining Fake News Debunking Decisions

Our findings corroborate the framing effect, a previously under-investigated aspect in explaining an AI system's decision to debunk fake news. Moreover, in line with the principle of compatibility (Proctor and Reeve 1990; Shafir 1993), our study results suggest that the explanations under the negative framing (i.e., "more false" and red color) are more compatible with the AI system's decision to "dispute" the news claim than the explanations under the positive framing (i.e., "more credible" and blue color). Thus, the explanations might have been more intuitive for the participants to interpret when debunking misinformation. Such results are also in agreement with prior research showing that users are likely to rely more on negative information than positive information to reject apps (Choe et al. 2013; Chen et al. 2015).

Moreover, our results point out the framing effect might be factor-dependent. We deployed three human understandable and interpretable factors in the bar chart. The results of Experiment 2 showed that the effect was only evident for contents and sources but not user comments. One possible reason is that sources of user comments, which are critical for judging misinformation (Seo et al. 2022), were missing. We did not obtain such results in Experiment 3, possibly due to the varied reliabilities and the low accuracy rating for the fake news. Future work can further test the observation by manipulating the sources of user comments.

Theories of human memory in cognitive psychology have informed our understanding of human susceptibility to misinformation (Loftus 2005; Pennycook, Cannon, and Rand 2018; Lee et al. 2023). The current study sheds light on extra cognitive factors that can be considered and further explored in future XAI work for misinformation mitigation.

The Impacts of System Reliability

With the implied truth effect (Pennycook et al. 2020), it is expected that participants should have little doubt in judging real news when fake news warnings are absent. Opposite to the prediction, our studies showed that participants did not credit real news cases by default but had concerns about *miss* errors (i.e., false negatives) of the AI system regardless of the framing. Also, we included two pieces of real news from CNN, which has been viewed as a source of fake news (Mastrine 2024). Thus, the extra explanations seemed to have helped participants detect some "miss errors" and made them biased toward judging more news as fake. Such results are consistent with prior work, which showed that when automation systems make miss errors, users reduce their reliance on the system (Dixon, Wickens, and McCarley 2007; Rice 2009). Reliance refers to the status in which users refrain from a response when the system is silent or indicating normal operation (Chancey et al. 2017). When participants were informed of the increased system reliability, their criterion of judging a piece of news as fake was adjusted. Thus, the accuracy rating of real news was increased and no differences were observed across the conditions.

We also obtained a somewhat floor effect (StatisticsHowTo 2023) on the fake news accuracy rating in Experiment 3 compared to Experiments 1 and 2. One possible reason might be due to using MTurk toolkit provided by CloudResearch in Experiment 3, which could have excluded inattentive workers and enhanced data quality (Hauser et al. 2022). Another possible explanation is that participants might have been able to detect the fake news without any warning or explanation since the news set was collected in late 2021 while Experiment 3 was conducted in 2022.

Moreover, participants in the *CON* condition reduced their accuracy rating of fake news when the AI system became more reliable. Such higher *compliance* (i.e., users respond when a signal is issued) suggests that participants have concerns about false alarm (i.e., false positive) when the AI system was less reliable. However, we did not obtain such results in the two explanation conditions, suggesting that participants in those conditions did not vary their dependence on the AI system for the fake news evaluation. During design, we arbitrarily assigned score values to each factor in the bar chart. Participants might have questions about the quality of the AI explanations across the different fake claims (e.g., a score of "1" for Twitter in one trial and a score of "5" in another trial). Future work could better control the accountability of the AI explanations.

It is noteworthy that AI systems are not always reliable, showing errors of false alarm (i.e., false positives) or miss (i.e., false negatives). Our findings highlight the importance of informing users of different errors made by AI systems (Kocielnik, Amershi, and Bennett 2019) and investigating the interaction between AI system reliability and error type.

Multidimensional Trust in AI Systems

A higher trust score was consistently obtained in the warning only condition, which are in agreement with findings in previous studies (Seo, Xiong, and Lee 2019; Epstein et al.

2022). Such results can be understood by the effect of *familiarity* on trust. Literature in different fields has shown that familiarity contributes to building trust (Gulati 1995; Barr 1999; Zhang, Ghorbani et al. 2004; Gulati and Sytch 2008). While the proposed explanations in our study were novel to the participants, they could be familiar with the warning label, which was similar to that deployed on Facebook. Moreover, the warning icon and red color have been widely used in everyday life to indicate risks or hazards (Wogalter, DeJoy, and Laughery 1999). Thus, even though participants' accuracy decisions could be influenced by the extra explanations, they still showed more trust in the warning itself.

System performance has been proposed as one of the bases for human-automation trust (Lee and See 2004). We explicitly varied the AI system's performance along two levels in the descriptions, and a degraded trust was evident when the AI system became unreliable. Yet, we did not obtain any interaction between warning and reliability, suggesting that human trust in AI systems is multidimensional.

Limitations

There are several limitations in the current study. *First*, we chose to recruit MTurk workers for a large sample. Although MTurk workers' demographics are more diverse compared to college students' (Weigold and Weigold 2022), they do not represent the whole population (Burnham, Le, and Piedmont 2018). Future studies could consider more comprehensive recruiting methods. *Second*, it can be difficult for platforms to disclose the reliability of their misinformation-detecting systems. Yet, social media platforms such as Facebook and Twitter have been actively responding to mitigate fake news and are well aware of the issue of information transparency.[5] Thus, if our findings could be continuously verified through follow-up studies, there is a good chance that those platforms will take the initiative to introduce warnings with explanations and provide reliability information to online users. *Third*, we only varied the system reliability in the description but did not answer the question: How does the reliability information conveyed through experience influence human trust in the AI system? Future work can test the possible description and experience gap (Hertwig et al. 2004). *Lastly*, our explanations could be unfamiliar or not intuitive for some participants. Thus, if some designers consider creating a warning with explanations, then they may want to consider users' graph literacy and highlight the contents' credibility information, which was considered the most for participants' perceived accuracy rating.

Conclusion

In order to verify the effect of a *warning with explanations*, we conducted three experiments. We found that *the effect of a warning with explanations on participants' perceived accuracy rating depends on the reliability of the AI system*. If the reliability information is unknown, the negatively-framed warning with explanations is more influential to participants, as they did not trust the system. When the reliability of the system was known to be high, a warning with explanations was not in effect, rather only a warning message was effective. Accordingly, if the level of reliability of the fake news detection system cannot be revealed, providing a warning with negatively framed explanations showing how the AI system evaluated news veracity can assist participants in avoiding fake news.

[5]https://transparency.fb.com/

Broader Impact and Ethical Statement

Our research protocol was approved by the Institutional Review Board (IRB) of the authors' institution. We asked for informed consent from each participant. We made sure to take suitable steps in our data collection and analysis to ensure an ethical study and preserve user privacy. In addition, we did not name any MTurk accounts in this paper to protect participants' privacy. Moreover, we note that we did not debrief the participants. Although we labeled warnings on all fake news in the experiments, we acknowledge that the lack of debriefing in our experiments could have potentially harmful effects on some participants. However, the prior studies showed that misinformation studies did not significantly increase participants' long-term susceptibility to misinformation used in the experiments (Murphy et al. 2020). With the development of AI, the expectation of trustworthy AI has increased. Transparency is an important factor for trustworthy AI. Our work addresses AI transparency at the levels of the entire system and specific predictions. Our findings reveal the unintended negative consequences by focusing on specific predictions only. Thus, it is essential to explain how an AI system behaves in a particular case and how it functions in general.

Acknowledgements

We thank the anonymous reviewers for their constructive comments and suggestions. This research was supported in part by Penn State under the PSU SSRI Seed Grant and the National Science Foundation under grants 1820609, 1915801, and 2121097.

References

Adadi, A.; and Berrada, M. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6: 52138–52160.

Arrieta, A. B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58: 82–115.

Barr, A. 1999. *Familiarity and trust: An experimental investigation*. University of Oxford.

Bode, L.; and Vraga, E. K. 2018. See something, say something: Correction of global health misinformation on social media. *Health Communication*, 33(9): 1131–1140.

Bode, L.; and Vraga, E. K. 2021. Correction Experiences on Social Media During COVID-19. *Social Media + Society*, 7(2). https://doi.org/10.1177/20563051211008829.

Burnham, M. J.; Le, Y. K.; and Piedmont, R. L. 2018. Who is Mturk? Personal characteristics and sample consistency of these online workers. *Mental Health, Religion & Culture*, 21(9-10): 934–944.

Chancey, E. T.; Bliss, J. P.; Yamani, Y.; and Handley, H. A. 2017. Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human Factors*, 59(3): 333–345.

Chen, J.; Gates, C. S.; Li, N.; and Proctor, R. W. 2015. Influence of risk/safety information framing on android app-installation decisions. *Journal of Cognitive Engineering and Decision Making*, 9(2): 149–168.

Chen, J.; Mishler, S.; Hu, B.; Li, N.; and Proctor, R. W. 2018. The description-experience gap in the effect of warning reliability on user trust and performance in a phishing-detection context. *International Journal of Human-Computer Studies*, 119: 35–47.

Cheng, H.-F.; Wang, R.; Zhang, Z.; O'Connell, F.; Gray, T.; Harper, F. M.; and Zhu, H. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Choe, E. K.; Jung, J.; Lee, B.; and Fisher, K. 2013. Nudging people away from privacy-invasive mobile apps through visual framing. In *IFIP Conference on Human-Computer Interaction*, 74–91. Springer.

Clayton, K.; Blair, S.; Busam, J. A.; Forstner, S.; Glance, J.; Green, G.; Kawata, A.; Kovvuri, A.; Martin, J.; Morgan, E.; et al. 2020. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, 42(4): 1073–1095.

Cleveland, W. S. 1985. *The elements of graphing data*. Wadsworth Publ. Co.

Cui, L.; Shu, K.; Wang, S.; Lee, D.; and Liu, H. 2019. defend: A system for explainable fake news detection. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2961–2964.

Dai, S.-C.; Hsu, Y.-L.; Xiong, A.; and Ku, L.-W. 2022. Ask to know more: Generating counterfactual explanations for fake claims. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2800–2810.

Dixon, S. R.; Wickens, C. D.; and McCarley, J. S. 2007. On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors*, 49(4): 564–572.

Dzindolet, M. T.; Peterson, S. A.; Pomranky, R. A.; Pierce, L. G.; and Beck, H. P. 2003. The role of trust in automation reliance. *International Journal of Human-computer Studies*, 58(6): 697–718.

Epstein, Z.; Foppiani, N.; Hilgard, S.; Sharma, S.; Glassman, E.; and Rand, D. 2022. Do explanations increase the effectiveness of AI-crowd generated fake news warnings? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 183–193.

Flanagin, A. J.; and Metzger, M. J. 2000. Perceptions of Internet information credibility. *Journalism & Mass Communication Quarterly*, 77(3): 515–540.

Gaziano, C.; and McGrath, K. 1986. Measuring the concept of credibility. *Journalism Quarterly*, 63(3): 451–462.

Greene, C. M.; and Murphy, G. 2021. Quantifying the effects of fake news on behavior: Evidence from a study of COVID-19 misinformation. *Journal of Experimental Psychology: Applied*, 27(4): 773.

Gulati, R. 1995. Does familiarity breed trust? The implications of repeated ties for contractual choice in alliances. *Academy of Management Journal*, 38(1): 85–112.

Gulati, R.; and Sytch, M. 2008. Does familiarity breed trust? Revisiting the antecedents of trust. *Managerial and Decision Economics*, 29(2-3): 165–190.

Gunning, D.; and Aha, D. 2019. DARPA's explainable artificial intelligence (XAI) program. *AI Magazine*, 40(2): 44–58.

Guo, Z.; Schlichtkrull, M.; and Vlachos, A. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10: 178–206.

Hauser, D. J.; Moss, A. J.; Rosenzweig, C.; Jaffe, S. N.; Robinson, J.; and Litman, L. 2022. Evaluating CloudResearch's Approved Group as a solution for problematic data quality on MTurk. *Behavior Research Methods*, 1–12.

Hauser, D. J.; and Schwarz, N. 2016. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, 48(1): 400–407.

Hertwig, R.; Barron, G.; Weber, E. U.; and Erev, I. 2004. Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, 15(8): 534–539.

Hoff, K. A.; and Bashir, M. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3): 407–434.

Horne, B. D.; Nevo, D.; O'Donovan, J.; Cho, J.-H.; and Adalı, S. 2019. Rating reliability and bias in news articles: Does AI assistance help everyone? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 247–256.

Jia, C.; Boltz, A.; Zhang, A.; Chen, A.; and Lee, M. K. 2022. Understanding effects of algorithmic vs. community label on perceived accuracy of hyper-partisan misinformation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–27.

Kocielnik, R.; Amershi, S.; and Bennett, P. N. 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*, 1–14.

Kreps, S. E.; and Kriner, D. L. 2022. The COVID-19 infodemic and the efficacy of interventions intended to reduce misinformation. *Public Opinion Quarterly*, 86(1): 162–175.

Lazer, D. M.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; et al. 2018. The science of fake news. *Science*, 359(6380): 1094–1096.

Lee, J. D.; and See, K. A. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1): 50–80.

Lee, M. K. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1): 2053951718756684.

Lee, S.; Seo, H.; Lee, D.; and Xiong, A. 2023. Associative inference can increase people's susceptibility to misinformation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 530–541.

Lin, X.; Spence, P. R.; and Lachlan, K. A. 2016. Social media and credibility indicators: The effect of influence cues. *Computers in Human Behavior*, 63: 264–271.

Lipton, Z. C. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3): 31–57.

Litman, L.; Robinson, J.; and Abberbock, T. 2017. TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 49(2): 433–442.

Loftus, E. F. 2005. Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory*, 12(4): 361–366.

Lu, Z.; Li, P.; Wang, W.; and Yin, M. 2022. The Effects of AI-based Credibility Indicators on the Detection and Spread of Misinformation under Social Influence. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–27.

Martel, C.; and Rand, D. G. 2023. Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology*, 101710.

Mastrine, J. 2024. Is CNN Fake News? https://www.allsides.com/blog/cnn-fake-news.

Mayer, R. C.; Davis, J. H.; and Schoorman, F. D. 1995. An integrative model of organizational trust. *Academy of Management Review*, 20(3): 709–734.

Mohseni, S.; Yang, F.; Pentyala, S. K.; Du, M.; Liu, Y.; Lupfer, N.; Hu, X.; Ji, S.; and Ragan, E. D. 2021. Machine Learning Explanations to Prevent Overtrust in Fake News Detection. In *ICWSM*, 421–431.

Molina, M. D.; Sundar, S. S.; Le, T.; and Lee, D. 2021. "Fake news" is not simply false information: A concept explication and taxonomy of online content. *American Behavioral Scientist*, 65(2): 180–212.

Mosallanezhad, A.; Karami, M.; Shu, K.; Mancenido, M. V.; and Liu, H. 2022. Domain Adaptive Fake News Detection via Reinforcement Learning. In *Proceedings of the ACM Web Conference 2022*, 3632–3640.

Murphy, G.; Loftus, E.; Grady, R. H.; Levine, L. J.; and Greene, C. M. 2020. Fool me twice: How effective is debriefing in false memory studies? *Memory*, 28(7): 938–949.

Nguyen, A. T.; Kharosekar, A.; Krishnan, S.; Krishnan, S.; Tate, E.; Wallace, B. C.; and Lease, M. 2018. Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, 189–199.

Norman, G. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, 15(5): 625–632.

Peer, E.; Rothschild, D.; Gordon, A.; Evernden, Z.; and Damer, E. 2022. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 54(4): 1643–1662.

Pennycook, G.; Bear, A.; Collins, E. T.; and Rand, D. G. 2020. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, 66(11): 4944–4957.

Pennycook, G.; Cannon, T. D.; and Rand, D. G. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, 147(12): 1865–1880.

Pieters, W. 2011. Explanation and trust: what to tell the user in security and AI? *Ethics and Information Technology*, 13(1): 53–64.

Proctor, R. W.; and Reeve, T. G. 1990. Research on stimulus-response compatibility: Toward a comprehensive account. In *Advances in psychology*, volume 65, 483–494. Elsevier.

Reis, J. C.; Correia, A.; Murai, F.; Veloso, A.; and Benevenuto, F. 2019. Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2): 76–81.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Rice, S. 2009. Examining single-and multiple-process theories of trust in automation. *The Journal of General Pychology*, 136(3): 303–322.

Savolainen, R. 2021. Assessing the credibility of COVID-19 vaccine mis/disinformation in online discussion. *Journal of Information Science*, 01655515211040653.

Seo, H.; Xiong, A.; and Lee, D. 2019. Trust It or Not: Effects of Machine-Learning Warnings in Helping Individuals Mitigate Misinformation. In *Proceedings of the 10th ACM Conference on Web Science*, 265–274.

Seo, H.; Xiong, A.; Lee, S.; and Lee, D. 2021. (In)effectiveness of Accumulated Correction on COVID-19 Misinformation. In *TMS Proceedings 2021*. Https://tmb.apaopen.org/pub/ss8t2ayg.

Seo, H.; Xiong, A.; Lee, S.; and Lee, D. 2022. If You Have a Reliable Source, Say Something: Effects of Correction Comments on COVID-19 Misinformation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 896–907.

Shafir, E. 1993. Choosing versus rejecting: Why some options are both better and worse than others. *Memory & Cognition*, 21(4): 546–556.

Shu, K.; Cui, L.; Wang, S.; Lee, D.; and Liu, H. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 395–405.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1): 22–36.

StatisticsHowTo. 2023. Floor Effect / Basement Effect: Definition. https://www.statisticshowto.com/floor-effect/.

Toreini, E.; Aitken, M.; Coopamootoo, K.; Elliott, K.; Zelaya, C. G.; and Van Moorsel, A. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 272–283.

Tversky, A.; and Kahneman, D. 1981. The framing of decisions and the psychology of choice. *Science*, 211(4481): 453–458.

Tversky, A.; and Kahneman, D. 1986. Rational Choice and the Framing of Decisions. *Journal of Business*, S251–S278.

Wang, X.; and Yin, M. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, 318–328.

Weigold, A.; and Weigold, I. K. 2022. Traditional and modern convenience samples: An investigation of college student, Mechanical Turk, and Mechanical Turk college student samples. *Social Science Computer Review*, 40(5): 1302–1322.

Westerman, D.; Spence, P. R.; and Van Der Heide, B. 2014. Social media as information source: Recency of updates and credibility of information. *Journal of Computer-mediated Communication*, 19(2): 171–183.

Wogalter, M. S.; DeJoy, D.; and Laughery, K. R. 1999. *Warnings and risk communication*. CRC Press.

Wu, L.; Morstatter, F.; Carley, K. M.; and Liu, H. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter*, 21(2): 80–90.

Yaqub, W.; Kakhidze, O.; Brockman, M. L.; Memon, N.; and Patil, S. 2020. Effects of Credibility Indicators on Social Media News Sharing Intent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

Zhang, J.; Ghorbani, A. A.; et al. 2004. Familiarity and Trust: Measuring Familiarity with a Web Site. In *PST*, 23–28. Citeseer.

Zhou, X.; and Zafarani, R. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5): 1–40.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes
   (g) Did you discuss any potential misuse of your work? No, because the potential risk of misuse is minimal.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes
   (b) Have you provided justifications for all theoretical results? Yes
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes
   (e) Did you address potential biases or limitations in your theoretical framework? Yes
   (f) Have you related your theoretical results to the existing literature in social science? Yes
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? NA
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, see a URL in footnote 3 for new assets (participants exclusion details).
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? No, but we present critical instructions, stimuli, and questions in the paper.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes
   (d) Did you discuss how data is stored, shared, and deidentified? No, we discussed how data is stored, shared, and deidentified in the IRB protocol, but we did not discuss it in our paper.