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 apply 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 a warning with AI 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 systems. Adding the explanations can potentially increase participants' suspicions on miss errors (i.e., false negatives) of the AI systems. Furthermore, more trust is shown in the 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 (Ha, Andreu Perez, and Ray 2021), companies such as Meta and Twitter have created warning systems to debunk fake news.[2] Previous works (Pennycook, Cannon, and Rand 2018; Clayton et al. 2020) have shown that a debunking warning label plays an effective role in mitigating fake news.

Meanwhile, active efforts have 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; Mohseni et al. 2021).

The value of the 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; Lu et al. 2022). For example, Lu et al. showed that presenting AI-based credibility indicators is effective in nudging participants into aligning their misinformation detection with the AI model's prediction. Seo et al. found that ML warning with explanation increased participants' ability to detect fake news more than the warning only when a news source was not provided.

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, 1985). Moreover, prior work (Seo, Xiong, and Lee 2019) presented that although AI explanations help people detect misinformation better, participants' trust in the AI system decreased. To fill the gap, in this work, we investigate whether explaining how an AI system debunks misinformation can improve the effectiveness of misinformation warnings. We focus on COVID-19 fake news considering its timely importance. Specifically, we examine the following research questions(RQs).

- **RQ 1.** Will a misinformation warning with explanations improve humans' ability to detect fake news compared to the warning only? If so, will a positive framing work better than a negative framing for the explanations?
- **RQ 2.** Will the effect of misinformation warning with explanations depend on the AI system's reliability?
- **RQ 3.** Will the misinformation warning with explanations increase humans' 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 humans' detection of mis-

[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

information with 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 AI system's reliability (i.e., whether the AI systems will make a lot of mistakes) (**RQ2**). We also evaluated humans' trust in AI systems (**RQ3**).

The results of our experiments suggest that the credibility explanations under negative framing (i.e., NEG) decrease humans' perceived accuracy ratings of fake news. Such results underline the necessity to consider the framing effect in examining the effectiveness of AI explanations in human misinformation detection. However, we find that humans 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 systems. 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 humans' misinformation detection can be influenced by explanations, they show more trust in the warning itself. Moreover, such trust increases when the AI system's reliability becomes higher. Our main contributions are summarized as follows.

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

Related Work

Misinformation and Correction

Rampant misinformation online has led many researchers (Bode and Vraga 2015; Pennycook, Cannon, and Rand 2018; Van der Meer and Jin 2020) and leading social media companies [3] to conceive effective ways to correct misinformation. The approaches to reducing misinformation online usually go through two steps (Seo et al. 2022). One is to detect which information is accurate through human fact-checkers (Lim 2018; Singer 2021) or machine learning (Shu et al. 2017). The second step is to correct the falsity of information using warning labels based on the evaluations (Vraga and Bode 2017; Seo, Xiong, and Lee 2019). Considering the roles of users as the main actors of information sharing (Bechmann and Lomborg 2013; Boyd and Ellison 2007), it is critical to discover effective ways to correct misinformation resulting in users' distrust in fake news.

Explainable Artificial Intelligence (XAI) and Misinformation Correction

Much recent work has been conducted to examine factors and designs enabling users to accept AI systems (Kim and Song 2020; Calisto, Nunes, and Nascimento 2022; Chong et al. 2022; Wintersberger et al. 2022). In line with those studies, XAI aims to make AI results more interpretable with explanations (Adadi and Berrada 2018; Gunning and Aha 2019; Arrieta et al. 2020). Many misinformation detection systems have been developed (Zhou and Zafarani 2020), but there is little research on how to display the detection results in a way that users can understand. Several related studies explored the effect of explanation in terms of AI evaluation on fake news (Mohseni et al. 2021) or investigated the effect of warning with a certain type of explanation (Horne et al. 2019; Epstein et al. 2022). However, those studies either did not explain concretely how an AI system derived each prediction (i.e., local) or how the AI system behaves in general (i.e., global). To fill the gap, our study explored the effect of a warning with an explanation showing factors that an AI system evaluates for fake news detection. Based on Seo et al. (Seo, Xiong, and Lee 2019) which confirmed an effect of a transparent machine learning warning, we designed our warning with an explanation to extend the research scope along with the framing effect theory and the impact of an AI system's reliability.

Credibility for Misinformation Correction

Even though it is hard to find universal agreement on the concept of credibility across different fields (Savolainen 2021), credibility has been used as a major 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 detecting misinformation. A number of studies have investigated the credibility of information on social media (Kang 2010; Kim 2010; Westerman, Spence, and Van Der Heide 2014; Lin, Spence, and Lachlan 2016; Savolainen 2021). Among them, Savolainen et al. suggested a conceptual framework of credibility by dividing two approaches including the credibility of the author and the credibility of mis/disinformation content. Meanwhile, Molina et al. (Molina et al. 2021) organized features of real news and fake news which are used for evaluation of news veracity. Based on these studies, we created our credibility factors for an explanation.

Framing Effects

Explaining the basis of the AI system's judgment is ultimately to help users understand the explanation effectively to increase the user's trust in the system (Pieters 2011). In order for users to be receptive to explanations, an effective explanation needs to be provided (Gunning and Aha 2019). Charts are often used to explain effectively the determinants of an AI system (Cheng et al. 2019; Wang and Yin 2021), which leads to the question of how to visualize charts more effectively. One approach could be to consider different framing options. Tversky and Kahneman addressed the framing effect first by explaining that people's willingness in taking risks can depend on how options are

[3] https://www.facebook.com/help/1952307158131536

presented (Tversky and Kahneman 1981, 1985). In the privacy domain, Choe et al. investigated the effect of the privacy rating in the context of the framing effects through visual representations of an app to warn the level of the app's privacy protection. The study confirmed the effect of a positively framed rating icon. On the contrary, in the health communication domain, Rosenblatt et al. (Rosenblatt et al. 2018) found that negatively framed warnings are more effective than positively framed warnings. Greene et al. (Greene and Murphy 2021)'s study presented no effects of providing a general warning about the dangers of online misinformation regardless of framing. Despite the different targets and framing designs, these studies throw insights that can be applied to our study. From the framing effect point of view, we propose to compare the negative framing and positive framing of the chart explanation to supplement the warning message against fake news.

Importance of Reliability

Previous studies showed that the effect of warning did not guarantee trust in the warning system. One of the factors could be the reliability of the system. Reliability is regarded as one factor of trust (Dzindolet et al. 2003). Several researchers (Dzindolet et al. 2003; Chancey et al. 2017) demonstrated the impact of reliability information in building users' trust. Dzindolet et al. 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 than the low-reliability system. Furthermore, Kocielnik et al. (Kocielnik, Amershi, and Bennett 2019) explored the impacts of different types of errors an AI system makes. These studies show how users react differently depending on how trustworthy an AI system is. In this context, we investigated the impact of reliability information on a warning system by comparing the impact of the high-reliability system and that of the low-reliability system. To the best knowledge, our study is the first study to cover the reliability information of the system in terms of explanation-based-warnings.

Trust in the AI System

The concept of trust is defined in various fields in different ways (Mayer, Davis, and Schoorman 1995; Hoff and Bashir 2015; Chancey et al. 2017). For example, Mayer et al. defined trust as a willingness to accept vulnerability. Trust in information systems indicates self-assurance by assessment of risks and alternatives (Pieters 2011). Furthermore, trust is the one that can have an impact on reliance on automation (Lee and See 2004). Machine learning researchers also pay attention to the importance of trust which is linked to justification issues of models (Ribeiro, Singh, and Guestrin 2016; Lipton 2018; Toreini et al. 2020). However, building trust in algorithmic systems can be a challenging task (Lee 2018).

An explanation increases trust by contributing to the transparency of the AI systems (Lipton 2018). Seo et al. (Seo, Xiong, and Lee 2019) conducted user studies testing the effects of warnings with and without explanation. The study stated machine learning graph warning increased participants' sensitivity in distinguishing fake news from real

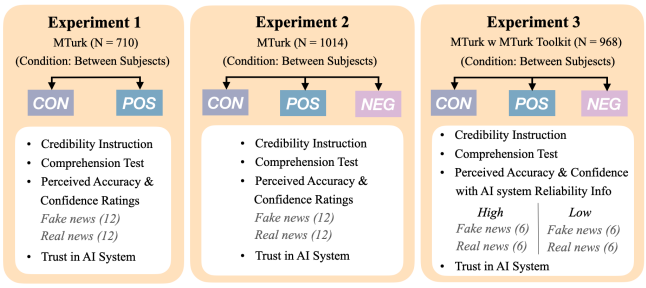


Figure 1: An overview of the experiment design. Experiments 1 and 2 focus on the framing effect. Experiment 3 focuses on reliability. *CON* means control, *POS* means positive framing and *NEG* means negative framing.

ones but did not increase trust in it. Epstein et al. (Epstein et al. 2022) replicated Seo et al.'s finding: trust did not increase with a warning with explanation although explanations increased the effect of a warning. XAI is developed to increase the users' trust in the system but warning studies have not demonstrated it so far. Therefore, our study tried to measure trust in the AI system after evaluating perceived accuracy ratings with warnings to see if our warning with explanations can have an impact on their trust.

The Present Study

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

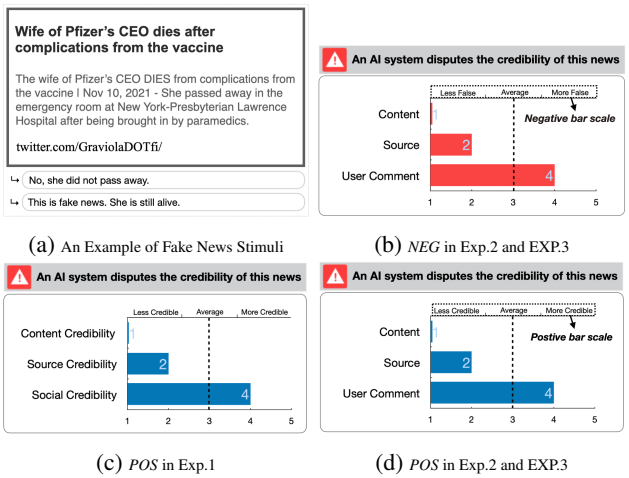
Participants

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%. MTurk workers were allowed to participate our study once.

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. After removing responses (1) submitted out of the U.S; (2) with duplicate IP; (3) failed the compre-

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%
	Prefer not to answer	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%
	60-69	4.8%	4.7%	6.6%
	70-79	0.3%	0.5%	1.4%
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%
	Prefer not to answer	0.3%	0.7%	0.7%
Education	High school	4.9%	7.4%	8.9%
	Some 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%
	Prefer not to answer	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.



(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: An example of fake news stimuli and the warning with explanations. A piece of COVID-19 fake news including the news title, a snippet of the news article, and a source was followed by two comments *(a)*. For fake news, a warning label was shown below the two comments. The bar chart for an explanation was shown below the warning label (*(b)-(d)*). For the warning-only condition (*CON*), only the warning message was shown. We modified the y-axis' names to minimize confusion in EXP.2 and EXP.3.

hension test; (4) failed the attention check and (5) completion time shorter than 3 min (average median completion time: 15 mins); the number of participants we accepted was 710, 1014, and 968, respectively. The high exclusion rate in Experiments 1 and 2 was to ensure our data quality,[4] which was necessary given the concerns on MTurk platform (Peer et al. 2022). In Experiment 3, we used the MTurk toolkit CloudResearch provides to exclude low-quality workers automatically (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 Articles. 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).

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

In our design, we assumed that each piece of fake news had been detected by an Artificial Intelligence (AI) system. A warning label was also shown for each piece of fake news, in which we made it clear that the fake news was disputed by an AI system (see Figure 2). There was a baseline condition in each experiment, in which we presented the warning label. Considering the robust effect of warning labels (Clayton et al. 2020) and our main interest in the effect of AI explanations, we omitted a condition without warning and defined the baseline with a warning label as the control (CON).

An abstract way of presenting the factors that AI systems considered when debunking fake news has been developed, which serves as an explanation for AI decision-making. Seo et al.(Seo, Xiong, and Lee 2019) presented that participants incorporated information provided by the summary index (i.e., bar chart) into their fake news evaluation. We adapted their design and created two types of explanations, positive framing (*POS*) and negative framing (*NEG*). As shown in Figure 2, we added a bar chart below the warning tag to illustrate the fake news credibility (*POS*) or falsity (*NEG*).

Positive Framing (*POS*) In the explanation interface, we present three factors that our hypothetical AI system considers, including content credibility based on the news title and news contents, source credibility based on the news source, and social credibility based on users' comments. A filled blue bar graph is accompanying each factor, and the length of each bar indicates credibility score that the AI

[4] See Exclusion Details in the supplementary materials.

system derived for the evaluation of 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. For score 1, a small blue bar is displayed in the bar graph. For score 5, the bar is displayed to the most right edge in the bar graph. For the bar graphs, an outline frame indicates the possible maximum score 5, so that the range of the score was clear to the participants and easy visual comparison was enabled among the bars (Cleveland 1985).

We made 12 bar charts to be added to the 12 pieces of fake news. We scored each factor using a 5-point score. We used either 1, 2, 4 or 1, 2, 5 for value combination avoiding 3, a neutral number. Each factor showed its high credibility (i.e., 4 or 5) four times among the 12 pieces of fake news. Moreover, we separated the 12 pieces of fake news into three sets and implemented a Latin-square design to counterbalance the credibility value combinations across the different sets. We focused on the credibility/falsity value combinations but did not control the value alignment for each factor. Our post-hoc analysis on the perceived accuracy rating of fake news showed no significant differences between the high and low source-credibility scores, suggesting limited impact.

Negative Framing (*NEG*) In addition to the positive framing using blue bars, we proposed a negative framing explanation (Choe et al. 2013). The interface was the same as the credibility explanation except that we changed 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 applied the same score set to the falsity explanation. Same as the credibility score, each factor had the highest value four times for the falsity score. Thus, compared to the credibility explanations with positive framing, fake news in the credibility explanations with negative framing was less false (See the *panels (b)* and *(d)* in Figure 2). To make the interface comparable to that of *NEG* in Experiment 2, we removed "Credibility" in the factor description of *POS* in Experiment 2 because it duplicates with the gauge on the top.

Procedure

Qualtrics was used for designing our online studies. After informed consent, participants were randomly assigned to one condition in each experiment. We first described our simulated warning system and asked participants questions to check their comprehension of our design. We asked two common questions for all conditions but added three more questions for *POS* and *NEG* to check participants' comprehension of each bar chart category. Then, the twenty-five pieces of stimuli were presented in a randomized order. Twelve of them included fake news and thirteen 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 instruction, their survey was terminated immediately. We paid those participants a base payment of $0.5.

We asked two questions to investigate participants' acceptance of the "claim" in the news article of each stimulus. First, participants rated their perceived accuracy rating of the news claim, "How accurate is the claim in the above news (1: very inaccurate, 7: very accurate)?". Then they answered a question about their confidence in their perceived accuracy rating, "How confident are you in answering the question above (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.", "I would like to see the AI system implemented on social media." Participants rated their agreement for 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

Our statistical analysis focused on three measures (*perceived accuracy rating*, *confidence in accuracy rating*, *trust in the AI system*). We manipulated two factors in Experiments 1 and 2, with *condition* between subjects and *news veracity* within subjects. In Experiment 3, we varied *reliability* as another within-subjects factor. We used SPSS version 29 for the data analysis.

Experiment 1

To quantify the effects, perceived accuracy and confidence results were entered into 2 (*news veracity*: real, fake) $\times$ 2 (*condition*: CON, *POS*) mixed analysis of variances (ANOVAs) with a significance level of .05. We chose ANOVAs since it is robust to yield the right answer even when distributional assumptions are violated (Norman 2010). Post hoc tests with Bonferroni correction were performed, testing all pairwise comparisons with corrected p values for possible inflation. The number of participants included for data analysis are: 413 (CON) and 297 (*POS*).

Perceived Accuracy Rating. As shown in Figure 3, participants clearly distinguished the real news (5.48) from the fake news (3.65), $F_{(1,708)} = 560.03$, $p < .001$, $\eta_p^2 = .442$. Regardless of the news veracity, participants in the *POS* gave lower ratings (4.47) than those in the CON (4.67), in general, $F_{(1,708)} = 6.74$, $p < .010$, $\eta_p^2 = .009$. Follow-up tests on each veracity revealed that the main effect of the condition was significant for the real news, $F_{(1,709)} = 6.35$, $p < .012$, $\eta_p^2 = .012$. but not the fake news, $F_{(1,709)} = 3.07$, $p = .080$, $\eta_p^2 = .004$. Nevertheless, the two-way interaction of *news veracity* $\times$ *condition* was not significant, $F < 1.0$.

Confidence in Accuracy Rating. Participants were confident in their perceived accuracy in general (average rating above 4, see Figure 3). Neither the main effect of *news veracity*, $F_{(1,708)} = 2.45$, $p = .118$, $\eta_p^2 = .003$, nor the main

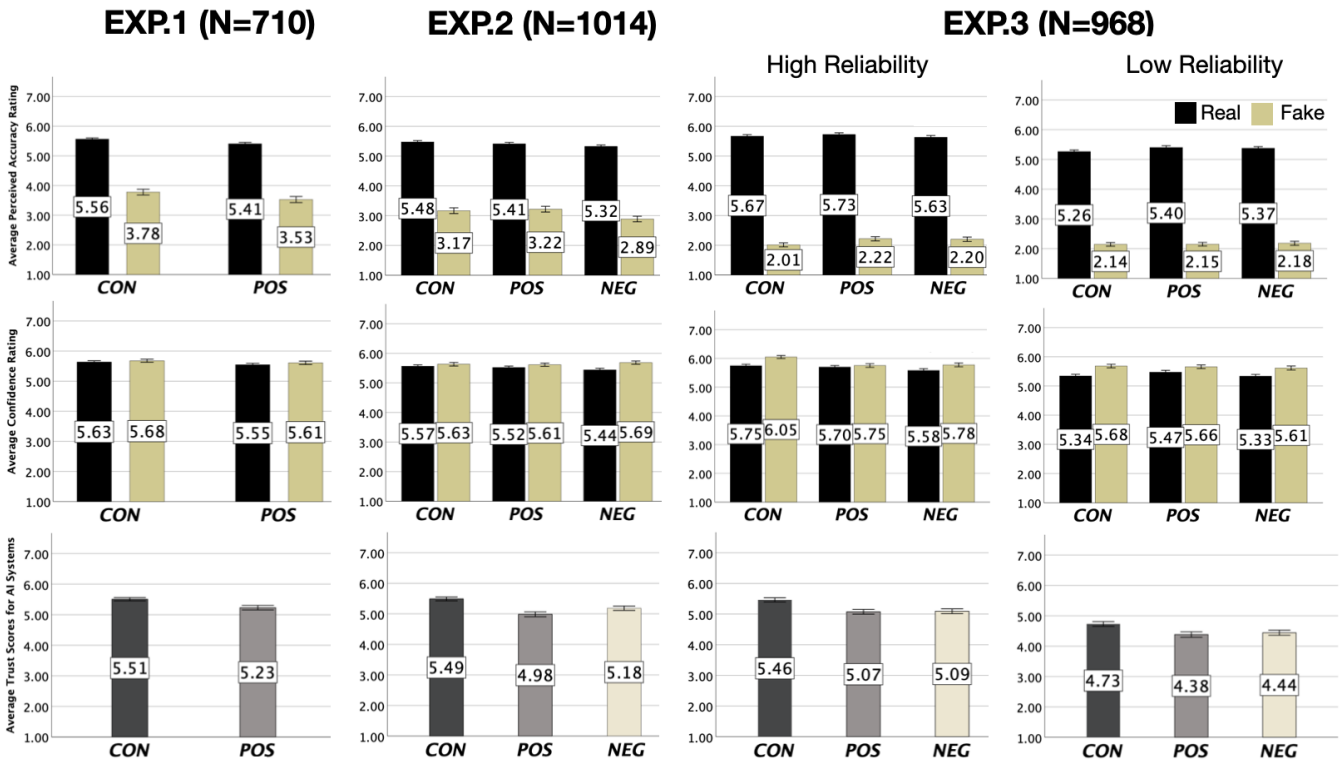


Figure 3: Mean values of perceived accuracy ratings (top row), confidence in the perceived accuracy decisions (middle row), and trust in the AI system (bottom row) across each condition in the three experiments (left column: Experiment (EXP.) 1, middle column: EXP.2, right column: EXP.3). *CON*: control condition showing only a warning label, *POS*: condition showing warning with explanations in a positive framing, *NEG*: condition showing warning with explanations in a negative framing.

effect of *condition*, $F_{(1,708)} = 1.52$, $p = .219$, $\eta_p^2 = .002$, were significant. The participants gave similar confidence ratings in their decisions of fake news (5.64) and real news (5.59). They also showed high confidence in both conditions (CON: 5.67 and *POS*: 5.58). The two-way interaction of *news veracity* $\times$ *condition* was not significant either, $F < 1.0$. Thus, the extra explanation in *POS* did not impact participants' confidence in their perceived accuracy ratings.

Trust in the AI System. We calculated the average ratings of the four questions asking about participants' trust in the AI system. Participants showed higher trust of the AI system in CON condition (5.51) than that in the *POS* condition (5.23), $F_{(1,709)} = 9.67$, $p = .002$, $\eta_p^2 = .013$.

Influential Credibility. To understand the relative weighting of the three factors in the *POS* condition, we analyzed the perceived accuracy rating results using ANOVA. The mean values of content credibility (3.59), source credibility (3.51), social credibility (3.48) showed no statistical difference, $F_{(1,296)} = 2.82$, $p = .094$, $\eta_p^2 = .009$.

Summary. In Experiment 1, we examined the effect of a warning label with explanations (i.e., a summary index showing the factors that an AI system considers when debunking fake news) on participants' perceived accuracy rating of fake claims. We observed that the participants who were exposed to the extra explanations (*POS*) decreased their perceived accuracy ratings on both fake and real news, particularly on real news. However, the participants showed similar confidence in their accuracy ratings across the conditions. Moreover, the participants showed more trust in the warning label (CON) than the label with extra explanations (*POS*). Thus, besides reducing participants' perceived accuracy of fake news, the extra explanations in *POS* have made participants more cautious overall.

Experiment 2

Human decision-making in risky contexts is influenced by how a problem is framed (The framing effect (Tversky and Kahneman 1981)). Considerations of compatibility indicates 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). Prior work on app selection has shown that the safety framing of the information is more compatible with a a selection task for best apps (Chen et al. 2015). Considering the AI system in our study is mainly debunking fake news, we conjecture that the ineffectiveness of the extra explanations in *POS* could be due to the positive framing in the design. We gauged the credibility of each factor ("more" means "better") and presented the credibility score using a blue color.

In Experiment 2, we proposed a *NEG* condition to further understand the effect of warning label with extra explanations. The overall experimental setting was the similar to Ex-

periment 1 except that we added *NEG* as another condition. The *POS* and *NEG* conditions were the same except that the wordings for bar gauge and the color of bar chart were different (See Figures 1 and 2 for the details). The number of participants included for data analysis are as follows: 390 (CON), 309 (*POS*), 315 (*NEG*). To quantify the effects, perceived accuracy rating and confidence rating were entered into 2 (*news veracity*: real, fake) $\times$ 3 (*condition*: CON, *POS*, *NEG*) mixed analysis of variances (ANOVAs). Post hoc comparisons were conducted in the same way as Experiment 1.

Perceived Accuracy Rating. Average results of the real and fake news for each condition are shown in Figure 3. Same as Experiment 1, the main effects of *news veracity*, $F_{(1,1011)} = 1353.97$, $p < .001$, $\eta_p^2 = .573$, and *condition*, $F_{(1,1011)} = 5.25$, $p = .005$, $\eta_p^2 = .010$, were significant. Specifically, participants gave higher accuracy ratings for the real news (5.40) than from the fake news (3.09). The average rating of *NEG* (4.11) was smaller than those of CON (4.32, $p = .01$) and *POS* (4.31, $p = .02$), respectively. Follow-up tests on each veracity revealed that the simple main effect of condition was only significant for fake news (CON: 3.17, *POS*: 3.22, *NEG*: 2.89), $F_{(1,1011)} = 3.24$, $p = .039$, $\eta_p^2 = .006$, but not real news (CON: 5.48, *POS*: 5.41, *NEG*: 5.33), $F_{(1,1011)} = 2.73$, $p = .066$, $\eta_p^2 = .005$. Such results were opposite to what we obtained in Experiment 1, indicating the framing effect. Nevertheless, the two-way interaction of news *veracity* $\times$ *condition* was not significant either, $F_{(1,1011)} = 1.15$, $p = .316$, $\eta_p^2 = .002$. As shown in Figure 3, participants in the *NEG* also showed a non-significant trend of reducing their perceived accuracy for real news.

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

Trust in the AI System. Participants' trust score varied across conditions, $F_{(1,1011)} = 13.54$, $p < .001$, $\eta_p^2 = .026$. Post-hoc pairwise comparisons revealed that the participants in the CON condition trust the AI system the most (5.49), which was significantly higher than those of *POS* (5.18, p=.006) and *NEG* (4.98, $p < .001$). However, participants' trust in the two explanation conditions were similar (p=.163).

Influential Credibility. We analyzed the mean values of the perceived accuracy rating among the three factors in *POS* and *NEG* conditions using mixed ANOVAs with 3 (*factor*: content, source, comments) as a within-subject factor and 2 (*condition*: *POS*, *NEG*) as a between-subject factor. The main effects of *factor*, $F_{(1,622)} = 4, 25$, $p = .04$, $\eta_p^2 = .007$, *condition*, $F_{(1,622)} = 6.02$, $p = .01$, $\eta_p^2 = .01$, and their two-way interaction, $F_{(1,622)} = 26.76$, $p < .001$, $\eta_p^2 = .04$, were all significant. Post hoc pairwise comparisons revealed significant differences among each pair for the *NEG* condition ($p_s \leq .027$). However, only the difference between content credibility (3.31) and comments credibility (3.13) was significant ($p = .009$) in the *POS* condition. Thus, among the three factors, participants mainly relied on the content for the veracity evaluation.

Summary. In Experiment 2, we further examined the effects of warning with explanations using negative framing (*NEG*) and positive framing (*POS*). The framing effect was revealed in the perceived accuracy rating of fake news and confidence ratings. Participants in the *NEG* tended to give lower accuracy rating for fake news and tended to be more confident in their ratings. Same as Experiment 1, participants were suspicious about the veracity of real claims, especially in the *NEG*.

Experiment 3

Across Experiments 1 and 2, we obtained that the participants did not always depend upon the warning or warning with explanations for their accuracy ratings. Especially, the participants were suspicious when the AI system did not tag a piece of real news (i.e., a system error of miss). Such results suggest that *local* explanations of specific decisions are not sufficient. Participants have concerns about the AI system's performance at a *global* level. Thus, in Experiment 3, we varied the reliability of the AI systems to detect fake news on two levels (high vs. low) 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 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 news and six pieces of fake news were shown, 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.

Perceived Accuracy Rating. Results of the average perceived accuracy ratings are shown in Figure 3. We ran mixed ANOVAs with 3 (*conditions*: CON, *POS*, *NEG*) $\times$ 2 (*news veracity*: real, fake) $\times$ 2 (*reliability*: low, high). Same as the prior two experiments, participants clearly distinguished

real news (5.51) from fake news (2.15), $F_{(1,965)} = 4021.36$, $p < .001$, $\eta_p^2 = .81$. However, the perceived accuracy rating of fake news in Experiment 3 was much lower than those in the prior two experiments (see Figure 3), indicating a floor effect (Cavanagh 2017).

The main effect of *reliability* was also significant, $F_{(1,965)} = 56.92$, $p < .001$, $\eta_p^2 = .056$, as well as the two-way interaction of *news veracity* × *reliability*, $F_{(1,965)} = 67.57$, $p < .001$, $\eta_p^2 = .065$. 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* × *reliability* × *condition* was also significant, $F_{(1,965)} = 5.72$, $p = .003$, $\eta_p^2 = .012$. For the real news, the increased accuracy ratings due to increased system reliability was similar across conditions ($p_s < .001$). Such results indicate that the system's reliability *did* address the participants' suspicions on the real news evaluation. However, for the fake news, the participants in the CON gave lower perceived accuracy rating when the AI system's reliability became higher ($p = .008$) but not those in the other two conditions ($p_s < .204$).

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 two-way interaction of *news veracity* × *condition* was significant, $F_{(1,965)} = 4.91$, $p = .008$, $\eta_p^2 = .010$. The participants showed higher confidence in their ratings of the fake news (5.76) than the real news (5.53). Such gap was more evident in the CON ($p < .001$) and *NEG* ($p < .001$) conditions than in the *POS* ($p = .013$) condition.

The main effect of *reliability*, $F_{(1,965)} = 5.72$, $p = .003$, $\eta_p^2 = .012$, the two-way interactions of *reliability* × *condition*, $F_{(1,965)} = 10.87$, $p < .001$, $\eta_p^2 = .022$, and *reliability* × *veracity*, $F_{(1,965)} = 7.21$, $p = .007$, $\eta_p^2 = .007$, were also significant. Participants were more confident in their accuracy ratings when the AI system's reliability become higher. For AI system with high reliability, participants' confidence ratings varied across the conditions. Specifically, participants' rating in the CON (5.90) was higher than that of *NEG* (5.68, p=.007). The confidence rating in the *POS* (5.73) showed no significant differences compared to the other two conditions ($p_s > .051$). For AI system with low reliability, participants' confidence ratings were similar across the conditions ($p_s > .797$). The confidence rating gap between real and fake news was larger when the system reliability is low (0.27) than when it is high (0.18).

Trust in the AI System. Participants' trust score varied across conditions, $F_{(1,965)} = 8.41$, $p < .001$, $\eta_p^2 = .017$. Post-hoc pairwise comparison showed that the participants trust the CON condition the most (5.10), followed by *POS* (5.07, $p = .003$) and *NEG* (4.77, $p < .001$). The main effect of *reliability* was also significant, $F_{(1,965)} = 331.51$, $p < .001$, $\eta_p^2 = .256$. Participants gave higher trust score for the system of high reliability (5.21) than that of low reliability (4.52). However, the two-way interaction of *reliability* × *condition* was not significant, $F < 1.0$.

Summary. We confirmed the system reliability is critical to address the participants' suspicious about the accuracy rating of real news. When the AI system's reliability became higher, participants reduced their perceived accuracy ratings of fake news in the *CON*, but not the other two framing conditions. Such results indicate the impact of other factors, which we discuss in the next section.

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., whether or not the AI system will make a lot of mistakes).

We obtained the evidence of the framing effect: participants who were exposed to the credibility explanation under negative framing tended to give lower accuracy ratings for fake news and tended to be more confident in their accuracy decisions. Yet, the participants did not always depend on the warning or warning with explanations for detecting misinformation. In particular, the participants became suspicious when the AI system did not tag a piece of real news (i.e., concern about the system error due to *miss*). Moreover, we confirmed that the system's reliability is critical to address such suspicion of the participants.

The Framing Effect on Explaining Fake News Debunking Decision

We proposed positively- and negatively-framed bar charts (i.e., a warning with explanations) showing three credibility factors that an AI system considers in debunking fake news. We examined the effect of those explanations on participants' perceived accuracy rating of fake claims. When we presented the positively-framed explanation (i.e., *POS*) in Experiment 1, we did not obtain any significant decrease in participants' accuracy rating on fake news. Instead, participants showed more doubts about real news and reduced their perceived accuracy ratings. In Experiment 2, we presented both positively- and negatively-framed (i.e., *POS*, *NEG*) explanations. Compared with the warning-only condition (i.e., *CON*), we found that their accuracy rating of fake news was significantly reduced with the negatively-framed explanations.

One possible reason for such results is that the explanations under 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 positive framing (i.e., "more credible" and blue color). Consequently, those explanations might have been more intuitive for the participants to interpret. Those results are similar to prior research that shows that users are often likely to rely

more on negative information than positive information to reject apps (Choe et al. 2013; Chen et al. 2015).

To the best of our knowledge, this is the first work to investigate the framing effect in explaining an AI system's decision to debunk fake news. We suggest future work to further explore the framing effect of debunking messages/explanations on the associated decisions and the compatibility between them.

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. We varied the AI system's reliability to detect fake news in Experiment 3 and confirmed that system reliability is critical to address the participants' suspicions about the accuracy of real news.

One possibility is that participants in the *POS* and *NEG* conditions might have paid more attention to news claims after viewing the bar charts, especially the three credibility factors. Since we did not explain how the AI system evaluated each factor and derived the score, participants seemed to have increased their bias to judge news claims as fake when the warning label and explanation were not presented for the real news. Thus, regardless of the framing, the extra explanations seemed to have made participants more conservative in detecting fake news when the reliability of the AI system is unsure. Such results are consistent with prior work which shows 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). However, when participants were informed of the system reliability in Experiment 3, their criterion of judging a piece of news as fake was adjusted. Thus, no significant differences were obtained across the three conditions for the real news in each reliability level.

Moreover, participants in the CON reduced their perceived accuracy rating of fake news when the AI system became more reliable. One possible explanation is that participants might have been able to detect the fake news without any warning or explanation since the news set we implemented was collected in late 2021 while the experiment was conducted in 2022. We also obtained a somewhat floor effect on the fake news accuracy rating in Experiment 3 compared to Experiments 1 and 2. Moreover, 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 pieces of fake news (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 and further investigate the interaction between the reliability and the framing effect.

Trust in the Warning

We investigated participants' trust in the AI system after finishing the main tasks about perceived accuracy rating and confidence rating. Across all experiments, *CON* consistently exhibited higher trust score compared to *POS* and *NEG*, respectively. Such results are consistent with previous studies (Seo, Xiong, and Lee 2019; Epstein et al. 2022). For example, Seo et al. found that a *Fact-checking* warning was the most trusted condition across two experiments regardless of the most effective warning type. In their second experiment, although the *Machine-Learning-Graph* warning (which inspired our *POS* warning) was effective, the *Fact-checking* warning was still trusted the most by the participants.

These 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 warning explanations in our study were novel to the participants, they could be familiar with the warning label which originated from Facebook. Moreover, the warning icon and red color have been widely used in our daily life to indicate risks or hazards (Wogalter, DeJoy, and Laughery 1999). Therefore, even though participants' accuracy decisions could be influenced by the extra explanations, they still showed more trust in the warning itself.

Higher Confidence in Fake News

Participants were confident in their perceived accuracy ratings in general (average ratings above 5 points out of 7). Between real and fake news, participants showed more confidence in their decisions about fake news than those of real news, particularly in Experiments 2 and 3. Those results revealed that the participants responded to the warning or the warning with explanations when a piece of fake news was labeled. Such high *compliance* (i.e., users respond when a signal is issued) suggests that participants did not worry about any false alarm (i.e., false positives) of the AI system.

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 results suggest that users tend to think about false negatives rather than false positives when a warning or a warning with explanations is presented. These findings highlight the importance of informing users of the possible error types of AI systems (Kocielnik, Amershi, and Bennett 2019).

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 2021), they do not represent the whole population (Burnham, Le, and Piedmont 2018). Future studies could consider more comprehensive recruiting methods. *Second*, we observed a larger perceived accuracy gap between real news and fake news 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 effectively exclude inattentive workers and enhance data quality (Hauser

et al. 2022). Another possible reason could be the time gap between experiments. EXP.3 was launched about 8 months after EXP.2 due to a natural delay in the research process. Consequently, participants might have been aware of the news veracity before the study. *Third*, it can be difficult for platforms to disclose the reliability of their misinformation-detecting systems. However, 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 will be a good chance that those platforms will take the initiative in introducing warnings with explanations and providing reliability information to online users. *Lastly*, our warning with explanation could be unfamiliar or not intuitive for some participants. Therefore, 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, when 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 to avoid fake news.

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.

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

Acknowledgements

We thank the anonymous reviewers for their constructive comments and suggestions. This research was supported in part by Penn State under 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.

Bechmann, A.; and Lomborg, S. 2013. Mapping actor roles in social media: Different perspectives on value creation in theories of user participation. *New Media & Society*, 15(5): 765–781.

Bode, L.; and Vraga, E. K. 2015. In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, 65(4): 619–638.

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.

Boyd, D. M.; and Ellison, N. B. 2007. Social network sites: Definition, history, and scholarship. *Journal of Computer-mediated Communication*, 13(1): 210–230.

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.

Calisto, F. M.; Nunes, N.; and Nascimento, J. C. 2022. Modeling adoption of intelligent agents in medical imaging. *International Journal of Human-Computer Studies*, 168: 102922.

Cavanagh, M. 2017. Floor Effect / Basement Effect: Definition. Accessed: 2023-01-10.

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.

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.

Chong, L.; Zhang, G.; Goucher-Lambert, K.; Kotovsky, K.; and Cagan, J. 2022. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior*, 127: 107018.

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.

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.

Ha, L.; Andreu Perez, L.; and Ray, R. 2021. Mapping recent development in scholarship on fake news and misinformation, 2008 to 2017: Disciplinary contribution, topics, and impact. *American Behavioral Scientist*, 65(2): 290–315.

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.

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.

Kang, M. 2010. Measuring social media credibility: A study on a measure of blog credibility. *Institute for Public Relations*, 4(4): 59–68.

Kim, S. 2010. Questioners' credibility judgments of answers in a social question and answer site. *Information Research*, 15(2): 15–2.

Kim, T.; and Song, H. 2020. The Effect of Message Framing and Timing on the Acceptance of Artificial Intelligence's Suggestion. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8.

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.

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.

Lim, C. 2018. Checking how fact-checkers check. *Research & Politics*, 5(3): https://doi.org/10.1177/20531680187868.

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.

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.

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.

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.

Rosenblatt, D. H.; Bode, S.; Dixon, H.; Murawski, C.; Summerell, P.; Ng, A.; and Wakefield, M. 2018. Health warnings promote healthier dietary decision making: Effects of positive versus negative message framing and graphic versus text-based warnings. *Appetite*, 127: 280–288.

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.

Singer, J. B. 2021. Border patrol: The rise and role of fact-checkers and their challenge to journalists' normative boundaries. *Journalism*, 22(8): 1929–1946.

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. 1985. The framing of decisions and the psychology of choice. In *Behavioral Decision Making*, 25–41. Springer.

Van der Meer, T. G.; and Jin, Y. 2020. Seeking formula for misinformation treatment in public health crises: The effects of corrective information type and source. *Health Communication*, 35(5): 560–575.

Vraga, E. K.; and Bode, L. 2017. Using expert sources to correct health misinformation in social media. *Science Communication*, 39(5): 621–645.

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. 2021. Traditional and Modern Convenience Samples: An Investigation of College Student, Mechanical Turk, and Mechanical Turk College Student Samples. *Social Science Computer Review.*, https://doi.org/10.1177/08944393211006847.

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.

Wintersberger, P.; van Berkel, N.; Fereydooni, N.; Tag, B.; Glassman, E. L.; Buschek, D.; Blandford, A.; and Michahelles, F. 2022. Designing for Continuous Interaction with Artificial Intelligence Systems. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–4.

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

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.