Understanding the Effects of AI-based Credibility Indicators When People Are Influenced By Both Peers and Experts

Supplemental Material

ZHUORAN LU, Purdue University, USA
PATRICK LI, Stanford University, USA
WEILONG WANG, Purdue University, USA
MING YIN, Purdue University, USA

1 DATA AND ANALYSIS CODE

All of the data and analysis code will be published after the acceptance of the paper.

2 THE EFFECTS OF AI AS THE NUMBER OF THE LAYPEOPLE PEERS VARIES

We separate our experimental data into different subgroups based on the number of veracity judgements from laypeople peers that the subject saw (i.e., the length of the pre-expert sequence). Then, within each subgroup, we examine whether the presence of AI-based credibility indicators have significant impacts on people's detection and spread of misinformation.

Figure 1 shows the results for Experiment 1. Note that we do not make comparisons on truth discernment or sharing discernment, since these two dependent variables are defined on the subject level and may not be well defined for some subject (e.g., if for all the tasks where a subject saw 3 peer judgements, the news in those tasks was always real, then this subject's truth discernment and sharing discernment can not be defined for the "3 peer judgements" subgroup). From Figure 1, we find that the presence of AI-based credibility indicators consistently leads to a significant increase in people's veracity judgement accuracy, regardless of the number of peer judgements people have seen (3 peer judgements: $p < 0.001$, Cohen's $d = 0.30$; 5 peer judgements: $p < 0.001$, Cohen's $d = 0.35$; 7 peer judgements: $p < 0.001$, Cohen's $d = 0.41$). Furthermore, for all three subgroups with varying size of the laypeople crowd, we find the consistent trend that presenting AI-based credibility indicator seems to nudge people into sharing more real news and less fake news, although the differences are not always significant (sharing intention of real news: 3 peer judgements: $p > 0.05$, 5 peer judgements: $p = 0.023$, Cohen's $d = 0.23$, 7 peer judgements: $p > 0.05$; sharing intention of fake news: 3 peer judgements: $p > 0.05$; 5 peer judgements: $p > 0.05$; 7 peer judgements: $p = 0.030$, Cohen's $d = 0.40$).

Similarly, Figure 2 shows the results for Experiment 2. Again, the impacts of AI-based credibility indicators on people's accuracy in detecting misinformation is consistently observed within different subgroups (3 peer judgements: $p < 0.001$, Cohen's $d = 0.37$; 5 peer judgements: $p < 0.001$, Cohen's $d = 0.38$; 7 peer judgements: $p < 0.001$, Cohen's $d = 0.41$). Moreover, we find that in all three subgroups, providing AI-based credibility indicators does not increase people's intention in sharing real news ($p > 0.05$ for 3,5,7 peer judgments), but tends to decrease people's intention in sharing fake news (3 peer judgements: $p = 0.031$, Cohen's $d = 0.41$; 5 peer judgements: $p = 0.029$, Cohen's $d = 0.49$; 7 peer judgements: $p > 0.05$). This is in line with the overall effects of AI-based credibility indicators that we have observed in Experiment 2.

Authors' addresses: Zhuoran Lu, Purdue University, West Lafayette, USA, lu800@purdue.edu; Patrick Li, Stanford University, Palo Alto, USA, prli@stanford.edu; Weilong Wang, Purdue University, West Lafayette, USA, wang4167@purdue.edu; Ming Yin, Purdue University, West Lafayette, USA, mingyin@purdue.edu.

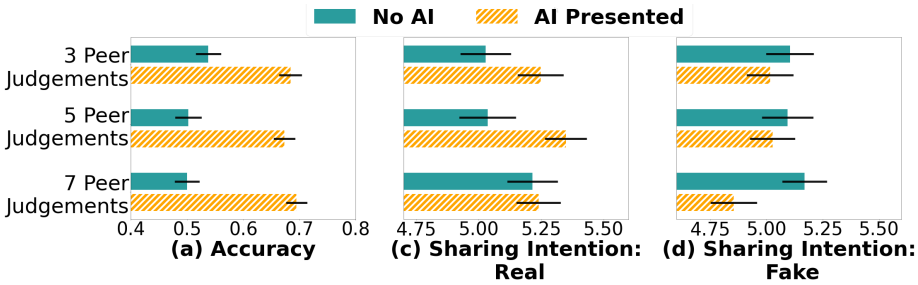

Fig. 1. The impacts of AI-based credibility indicators on subjects' ability in detecting misinformation and their intention to spread true and false information, when they are influenced by experts with self-claimed expertise (Experiment 1). Data is separated into three subgroups based on the number of veracity judgements made by laypeople peers in the pre-expert sequence. Error bars represent the standard errors of the mean.

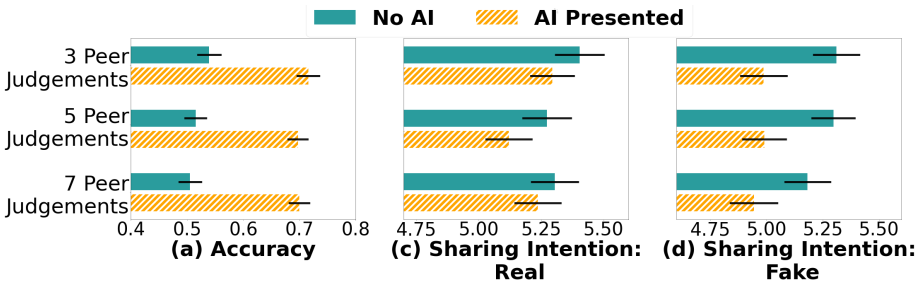

Fig. 2. The impacts of AI-based credibility indicators on subjects' ability in detecting misinformation and their intention to spread true and false information, when they are influenced by experts with verified expertise (Experiment 2). Data is separated into three subgroups based on the number of veracity judgements made by laypeople peers in the pre-expert sequence. Error bars represent the standard errors of the mean.

3 THE IMPACT OF EXPERT-AI AGREEMENT ON THE EFFECTS OF AI-BASED CREDIBILITY INDICATORS WHEN EXPERTS ARE VERIFIED

Dependent Var	d (Expert-AI Agree)	d (Expert-AI Disagree)	$\Delta\overline{d}$
Accuracy	0.31 [0.25, 0.38]	0.46 [0.39, 0.53]	-0.15***
Truth discernment	0.67 [0.44, 0.94]	0.97 [0.73, 1.22]	-0.30***
Sharing Intention: Fake	0.16 [0.05, 0.28]	0.18 [0.06, 0.30]	-0.02***

Table 1. Comparison of effect sizes (measured in Cohen's d and the 95% bootstrap confidence intervals) of the AI-based credibility indicators in the *Expert-AI agree* and *Expert-AI disagree* scenarios in Experiment 2. $\Delta\overline{d}$= d(Expert-AI agree) - d(Expert-AI disagree) is the difference of the average effect sizes. *** represents a significance level of 0.001.

To formally compare the effect sizes of AI-based credibility indicators between the scenarios where the experts agree or disagree with AI in Experiment 2, we again conduct bootstrap re-sampling ($K = 1000$) within each subgroup of data and estimated the effect size of AI-based credibility indicators using Cohen's d given each bootstrapped sample of the data. This time, we focus on estimating the sizes of the effects of AI-based credibility indicators on subjects' veracity judgement accuracy, truth discernment, and sharing intention on fake news, since these are the only three dependent variables for which at least marginal effects of AI were detected in both scenarios. The estimation results are shown in

Model/ Independent Var	**Model 1:** y = Sharing Intention			**Model 2:** y = Positive Peer Judgement			**Model 3:** y = Sharing Intention		
	Exp1	Exp2	Exp3	Exp1	Exp2	Exp3	Exp1	Exp2	Exp3
Intercept (C)	5.09^{***}	5.25^{***}	5.36^{***}	0.64^{***}	0.65^{***}	0.67^{***}	4.91^{***}	5.02^{***}	5.21^{***}
Positive Peer Judgement (β_1)							0.28^{**}	0.37^{***}	0.26^{**}
Expert Judgement - Real (β_2)	0.04	0.08	0.04				0.04	0.08	-0.005
Positive AI Prediction (β_3)	0.17^{*}	-0.08	0.13^{**}	0.11^{***}	0.11^{***}	0.11^{***}	0.14^{*}	-0.12	0.10^{*}
Negative AI Prediction (β_4)	-0.15^{*}	-0.32^{***}	-0.06	-0.17^{***}	-0.19^{***}	-0.15^{***}	-0.10	-0.25^{***}	-0.02

Table 2. Regressions for understanding how the impact of AI-based credibility indicators on people's sharing intention is mediated by the fraction of peers who make "positive" veracity judgements (i.e., consider the news to be *real*). * and *** represent significance levels of 0.05 and 0.001, respectively.

Table 1. As before, paired t-tests were used to examine whether differences in the mean values of the estimated effect sizes in the two scenarios are significantly different. Our results again suggest that when people are subject to social influence from both peers and experts with verified expertise, the effects of AI-based credibility indicators on people's detection and spread of misinformation are still more salient when the veracity judgements made by the verified experts are inconsistent with the AI model's predictions.

4 MEDIATION ANALYSIS ON PEOPLE'S SHARING INTENTION

We start by comparing whether AI-based credibility indicators or the experts have larger impacts on people's willingness to share news. To do so, we used linear regression models to predict participants' willingness to share news based on three factors: whether the expert's opinion indicated the news was real ("*Expert Judgment - Real*"), whether the AI-based credibility indicator suggested the news was real ("*Positive AI Prediction*"), and whether the AI suggested the news was fake ("*Negative AI Prediction*"). The regression results, shown in Table 2 (Model 1), reveal that across all three experiments, only AI-based predictions significantly influenced the participants' willingness to share the news. Specifically, positive AI predictions made participants more willing to share the news ($\beta_3 > 0$), while negative AI predictions reduced their willingness to share it ($\beta_4 < 0$). In contrast, expert judgment did not have a significant impact on participants' willingness to share.

We then explored how AI-based credibility indicators affect people's sharing intentions: Do they affect subjects directly or indirectly by influencing the veracity judgments of peers who reviewed the news beforehand? To investigate this, we conducted mediation analyses, reported in Table 2. For each subject, we first examined whether the AI-based credibility indicator's prediction on the news veracity would change the fraction of laypeople peers who would consider the news as real. The results in Model 2 of Table 2 show that in all three experiments, both positive and negative AI predictions significantly affected the fraction of preceding peers who would consider the news as real ($\beta_3 > 0, p < 0.001; \beta_4 < 0, p < 0.001$). Positive AI predictions increased the likelihood that peers considered the news to be real, while negative AI predictions increased the likelihood that peers considered the news to be fake.

In Model 3, we included the influence of all four parties (laypeople peers, expert, positive AI predictions, and negative AI predictions) to predict participants' sharing intentions. The results indicate that the fraction of positive peer judgments remained significantly influence participants' sharing intention ($\beta_1 > 0, p < 0.001$), and AI-based credibility indicators continued to influence sharing intentions to some extent—positive AI predictions increased the willingness of sharing ($\beta_3 > 0, p < 0.05$ for Experiment 1 and 3), while negative AI predictions decreased it ($\beta_4 < 0, p < 0.001$ for Experiment 2). This suggests that AI-based credibility indicators affect people's sharing intentions both directly and indirectly, through their impact on peers' veracity judgments, thus shaping the subjects' willingness to share news.