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

Zhuoran Lu
Purdue University
West Lafayette, Indiana, USA
lu800@purdue.edu

Patrick Li
Stanford University
Palo Alto, California, USA
prli@stanford.edu

Weilong Wang
Purdue University
West Lafayette, Indiana, USA
wang4167@purdue.edu

Ming Yin
Purdue University
West Lafayette, Indiana, USA
mingyin@purdue.edu

Abstract

In an era marked by rampant online misinformation, artificial intelligence (AI) technologies have emerged as tools to combat this issue. This paper examines the effects of AI-based credibility indicators in people's online information processing under the social influence from *both peers and* "*experts*". Via three pre-registered, randomized experiments, we confirm the effectiveness of accurate AI-based credibility indicators to enhance people's capability in judging information veracity and reduce their propensity to share false information, even under the influence from both laypeople peers and experts. Notably, these effects remain consistent regardless of whether experts' expertise is verified, with particularly significant impacts when AI predictions disagree with experts. However, the competence of AI moderates the effects, as incorrect predictions can mislead people. Furthermore, exploratory analyses suggest that under our experimental settings, the impact of the AI-based credibility indicator is larger than that of the expert's. Additionally, AI's influence on people is partially mediated through peer influence, although people automatically discount the opinions of their laypeople peers when seeing an agreement between AI and peers' opinions. We conclude by discussing the implications of utilizing AI to combat misinformation.

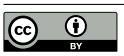
CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Machine learning**.

Keywords

misinformation, fake news, artificial intelligence, social influence, human-AI interaction

ACM Reference Format:
Zhuoran Lu, Patrick Li, Weilong Wang, and Ming Yin. 2025. Understanding the Effects of AI-based Credibility Indicators When People Are Influenced By Both Peers and Experts. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3706598.3713871



CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713871

1 Introduction

The widespread presence of online misinformation (i.e., "fake news") across social media has become a major challenge for citizens with many negative real-life consequences, from generating misconceptions to fueling distrust or even putting lives at risk [94, 114, 119, 129]. To help people tackle this challenge, many social media platforms have opted to provide their users with credibility indicators alongside news stories. The hope is that these indicators can help people better detect misinformation and therefore reduce their engagement with them. While social media platforms today largely utilize manual fact-checking to assess the credibility of different information [7, 18, 136], the rapid development of artificial intelligence (AI) technologies has shown the promise of scaling up this process by automating the assessment [40, 105, 122]. In fact, some platforms have already developed AI-based tools to signal potential misinformation to fact-checkers to speed up their work [3], and researchers have explored various designs to improve the accuracy of AI-based credibility indicators [30, 60, 67] and increase their impact on people [22, 27, 61].

Meanwhile, to understand the usefulness of the AI-based credibility indicators to *end users* on social media, it is critical to examine whether providing AI-based credibility indicators along with news can help users accurately evaluate the veracity of the news and dissuade them from sharing false information. A few recent studies have shown that when people consume information *independently*, the presence of AI-based credibility indicators can indeed improve their capability for differentiating true and false information and decrease their tendency to share fake news [42, 101, 134]. However, the real-world social media environment is much more complex. In reality, news often gets spread in the social network along different "paths." This means that when a user receives a piece of news, many other users on social media (i.e., the preceding users on the propagation path of the news) may have already read about the news. These preceding users may have even made some judgements about the credibility of the news that can be seen by the current user (e.g., via comments), which may influence their perceptions of and engagement with the news. To make this even more complicated, the influence of each of these preceding users may not be the same. For example, for news in domains that require specialized knowledge (e.g., health news), one may consider experts' judgements to be more trustworthy and thus be influenced by experts more than

their laypeople peers [76, 84, 99]. Thus, our first goal in this study is to answer the following question:

> **RQ1**: When people's perceptions of information credibility are influenced by *both peers and experts*, can the presence of AI-based credibility indicators help people better detect misinformation and reduce the spread of misinformation?

There are reasons to conjecture the answer either way. On the one hand, the AI-based credibility indicators, when accurate, may help reduce misinformation because they may trigger the *machine heuristic* among users [113]—Users may decide to align their belief in the veracity of the news with the AI model's predictions as they believe AI systems are "powerful" and will likely generate accurate predictions. If many of the preceding users do so due to their exposure to the AI-based credibility indicators, the positive effects of AI-based credibility indicators may even be amplified by the "*bandwagon effect*", i.e., people's tendency to follow what others think or do [58, 72, 78]. On the other hand, if users exhibit the "*authority heuristic*" [113] and mostly rely on the experts' opinions to form their own judgements, then the presence of AI-based credibility indicators may have limited impacts on them. It is also possible that users may start questioning the trustworthiness of AI-based credibility indicators after seeing some of their preceding users held beliefs about news veracity that were directly opposed to the AI's assessments; this again may imply minimal effects of the AI-based credibility indicators.

In addition, we speculate that when people are influenced by peers and experts, whether and to what extent AI-based credibility indicators can help reduce misinformation may be moderated by some contextual factors. This leads to our additional questions:

> **RQ2**: Does *the agreement between the expert and the AI-based credibility indicator* (i.e., whether they agree on the credibility of a piece of information) moderate the effects of AI-based credibility indicators?
> **RQ3**: Do the effects of AI-based credibility indicators change when the expert's expertise is *verified*, compared to when the expert's expertise is *self-claimed*?
> **RQ4**: How do the effects of AI-based credibility indicators change when the *competence* of the AI-based credibility indicator varies (e.g., when the AI-based credibility indicator has varying levels of accuracy)?

Again, it is difficult to make ex-ante predictions to these questions. For example, the effects of AI-based credibility indicators can be larger when the expert agrees with AI than when they disagree with each other, if people only consider AI predictions that are consistent with experts' judgements to be trustworthy. However, if people always believe in AI predictions more than the experts, we may arrive at the opposite conclusion. Similarly, the ways that people weigh the opinions of different parties (e.g., peers, experts, and AI) may significantly differ depending on whether the experts' expertise is verified or the competence level of AI, thus leading to challenges to answer RQ3 and RQ4, respectively.

Therefore, to answer these questions, we conducted a series of three pre-registered, randomized human-subject experiments on Amazon Mechanical Turk (MTurk). Participants in our experiments were recruited to review health-related news and determine their willingness to share them. To mimic how news gets diffused in social networks, when reviewing a piece of news, participants could also view the veracity judgements made by all preceding participants who had reviewed it. Moreover, among these preceding veracity judgements, we artificially inserted an "expert judgement". Specifically, in Experiment 1, we told participants that the expert's expertise in related domains (e.g., medicine, nursing, etc.) was claimed by themselves. Participants were randomly assigned to one of the two treatments—in the control ("No AI") treatment, participants did not have access to the AI-based credibility indicators when reviewing the news, while participants in the experimental ("AI presented") treatment had access. In the AI presented treatment, the credibility indicator was based on an AI model that was perfectly accurate in differentiating true and false information; this allowed us to examine the best possible effects that might be brought up by an "ideal" AI-based credibility indicator in reducing people's belief in and engagement with misinformation when they are subject to influences from both peers and experts.

To explore the potential moderating effects of contextual factors, we further conducted two replication studies of Experiment 1. In particular, in Experiment 2, we used the same perfectly accurate AI-based credibility indicator as that used in Experiment 1, but we informed participants that the experts' expertise had been verified. On the other hand, in Experiment 3, we followed the design of Experiment 1 to inform participants that experts' expertise was self-claimed. However, different from that in Experiment 1, in addition to the control "No AI" treatment, we created two experimental treatments involving the presence of AI-based credibility indicator by varying the competence level of AI, i.e., the "High Accuracy AI" and "Low Accuracy AI" treatments, and the accuracy of the AI-based credibility indicator was 80% and 55%, respectively, in these two treatments (see Table 1 for an overview of the designs of our three experiments).

Our experimental results show that when people are subject to both peer influence and expert influence, the presence of accurate AI-based credibility indicators can still significantly improve their ability to differentiate true information from false information and decrease their engagement with misinformation. This is true both when the expert's expertise is self-claimed and when it is verified, and the impacts of AI-based credibility indicators are larger when the experts' judgement and the AI prediction on news veracity do *not* align with one another. However, these positive effects of the AI-based credibility indicators heavily rely on the AI-based credibility indicators being accurate—as people lack the capability of differentiating the correctness of AI predictions, AI-based credibility indicators could also lead to undesirable effects on people's recognition of and engagement with misinformation when AI makes mistakes. Via a few exploratory analyses, we further reveal that the AI-based credibility indicators partially exert their effects on people via influencing their laypeople peers' veracity judgements. As such, people appear to slightly discount the opinions of their laypeople peers in determining the credibility of different information if they find the majority of their peers agree with the AI-based credibility indicator. It was also found that under our experimental settings, the total effects of AI-based credibility indicators on people's perceptions of and engagement with news, including their direct and indirect effects, are larger than those of

	Experiment 1	Experiment 2	Experiment 3
Treatment	No AI v.s. AI Presented	No AI v.s. AI Presented	No AI v.s. High Accuracy AI v.s. Low Accuracy AI
AI Accuracy	100%	100%	Low Accuracy AI: 55%, High Accuracy AI: 80%
Expert's Expertise	Self-claimed	Verified	Self-claimed
Targeted Research Questions	**RQ1, RQ2**	**RQ3**	**RQ4**

Table 1: Summary of the design of the three experiments.

the expert's. Together, these results highlight both the promise and potential limitations of leveraging AI-based credibility indicators to combat misinformation in real-world social media.

2 Related Work

2.1 Misinformation and Interventions

As an issue with far-reaching social impacts, misinformation has sparked great research interest, especially after the rise of social networks that make it easy to spread [1, 50, 126, 139]. Researchers have looked into the harms of misinformation [5, 6, 54, 91] and found its impact varies across contexts. For example, in low-stake scenarios such as entertainment [16, 68], misinformation is often used to manipulate perceptions or spark arguments. However, in high-stakes domains like politics [48, 83, 111] and health [112, 115, 127], misinformation poses more severe risks, potentially leading people to make inappropriate decisions with harmful consequences [79, 109]. Researchers also investigated the diffusion patterns of misinformation [11, 26, 36, 117, 123], and why people believe in and share misinformation [31, 39, 53, 89, 90, 121].

Recently, the issue of misinformation has become an increased concern, as the creation and spreading of misinformation becomes increasingly easy in the era of AI [43, 130]. In response, many interventions have been proposed to protect people from being misled by misinformation and reduce their engagement with misinformation. Early approaches focus on leveraging users' own capabilities to judge the credibility of information. For instance, accuracy prompting encourages users to critically think about the credibility of news stories before engaging with them [8, 47, 86, 87]. Another approach involves promoting strategies such as lateral reading, where users could verify information veracity by consulting multiple sources or searching online to make more informed judgments [45, 82]. Additional methods have also been developed to enhance users' capabilities in detecting potential misinformation. For instance, platforms may share expert consensus on a topic to "inoculate" the public against false information [10, 20, 120, 120], and systems have been developed to help center users' engagement with information around credibility [44]. These methods are relatively lightweight, though ultimately their effectiveness is based on the capability of users to learn to differentiate true and false information by themselves and make informed judgment.

Another more straightforward approach to assist people in combating misinformation is to conduct fact-checking and provide credibility signals along with the information. For example, social media platforms often signal the credibility of information to users through warning labels [13, 38, 49, 98]. It was found that these warning labels can often effectively reduce the perceived accuracy of misinformation by people and reduce people's intention to share misinformation [18, 71, 74, 103, 124, 134]. However, these warning labels are often produced based on manual fact-checking, relying on judgments of professional fact-checkers [35, 70, 73] or crowd-sourced annotations [34, 55, 88, 95, 118]. Although effective, this manual process is costly and struggles to scale with the rapid growth of content on online platforms.

2.2 AI-based Credibility Indicators

In recent years, extensive empirical research has explored whether and how AI model recommendations impact human decision-making in a wide range of scenarios [14, 15, 56, 59, 65, 66, 97, 135]. In the context of information spread, many efforts have been made to develop AI-based tools (e.g., machine learning models) to automatically evaluate the credibility of different information [21, 46, 52, 62, 63, 75, 92, 106, 131, 133]. The rise of generative AI has further accelerated this progress by improving the performance of classifiers [30, 60, 67], and enabling new interactive fact-checking systems [116, 132, 137]. These advancements bring about the possibility of providing real-time, AI-based credibility indicators to people as they process the information.

In some of the most recent experimental studies, it has been found that presenting AI models' predictions on the veracity of news to people can significantly increase people's ability to detect fake news and decrease their propensity to share fake news [42, 64, 80, 101, 134]. Meanwhile, the effectiveness of AI-based credibility indicators is also impacted by a wide range of factors [61]. For example, explanations are widely considered to enhance the effectiveness of AI-based credibility indicators by providing rationales for veracity judgments [22, 27]. Schmitt et al. [96] found that free text explanations could help improve non-experts' performance in detecting misinformation. On the other hand, Seo et al. [100] found that the framing of the explanations provided by AI-based credibility indicators moderates the effectiveness of the indicators. There are also studies that reveal null effects or time-varying effects of AI-based credibility indicators. For example, in some cases, it was found that AI-based warning labels on news headlines can not improve people's understanding of the veracity of the news, especially when they are convincingly wrong [23, 25, 107]. In addition, when personalized AI models are trained to be tailored to an individual's assessment of information credibility, the impact of personalized AI on people's credibility judgments was found to grow over time [46].

2.3 Complex Social Environment Where Information Spreads

Compared to the previous studies, our work focuses on understanding the effects of AI-based credibility indicators in a more

realistic social media environment where people are subject to *social influence from a crowd of mixed expertise*. During information diffusion, social influence may arise from various sources, including peers [17, 104], celebrities [81, 138], and other entities in the environment [4, 41]. It plays a crucial role in the formation of opinions, and may explain critical phenomena like herding and the bandwagon effect [32, 77]. As an example of how people's perception of and engagement with online information may be affected by social influence [2], a recent study revealed that after seeing other people criticize a news article as fake, users decreased their likelihood of sharing it [19]. On the other hand, seeing the engagement of others with a news post can increase the likelihood that a user will share and like the post [28]. Moreover, the influence brought up by each individual is not necessarily the same. For example, it was found that on Reddit, posts from users with more domain expertise (e.g., those who claimed to hold doctorate degrees) inspired more discussions and were perceived as more convincing by the readers of the posts [84], which may imply the "Halo effect" [24, 51]. On the conceptual level, Sundar [113] proposed the MAIN model, which suggests that people's determination of information credibility largely relies on various "heuristics"—what the laypeople peers believe, the experts state, or a machine (e.g., an AI model) predicts can all serve as the heuristics for people to use. While some recent research has started to examine how people judge information credibility in the presence of two heuristics (e.g., peers and experts, peers and AI) [9, 64, 125], in this study, we look into a more sophisticated setting in which three heuristics (i.e., peers, experts, and AI) might be presented simultaneously, and some of these heuristics may not be independent (e.g., AI can impact peer judgements), to understand the effects of AI-based credibility indicators.

3 Experiment 1: Effects of Perfect AI When People are Influenced by Peers and Self-Claimed Experts

In our first experiment, we set out to understand whether and how ideal AI-based credibility indicators (i.e., indicators that reach perfect accuracy) can help reduce misinformation, when people's processing of online information are influenced by both peers and experts with *self-claimed* expertise. We address this question by conducting a pre-registered, randomized human-subject experiment on Amazon Mechanical Turk (MTurk)[1].

3.1 Experimental Tasks

Participants were recruited to complete a series of tasks to review news headlines related to health, and then report their perceptions of and willingness to engage with them. In particular, we collected a dataset of 40 pieces of health-related news, which consisted of 20 true news headlines (i.e., "real news") and 20 false news headlines (i.e., "fake news"). We confirmed the veracity of each real news by cross-checking multiple reliable media sources or peer-reviewed publications. On the other hand, the fake news included in our dataset was previously disputed by authoritative sources or conflicted with verified information. We decided to use health-related news in our experiments since such news is prevalent in real-world social media, but judging its veracity can be challenging and may require a degree of domain expertise. Through a pilot study, we also found that people's independent accuracy in determining the veracity for each news in our dataset was mostly between 50% and 60%, suggesting that people have limited prior knowledge on such news. As a result, people may naturally be influenced by both their peers and the experts when evaluating the credibility of these health-related news headlines, which fits the purpose of our experiments well.

In each task, we randomly selected one piece of news from our dataset and presented it to the participant. The participant was asked to carefully review the news, as well as the opinions of those participants who reviewed it *before* them regarding the veracity of this news. Then, the participant made a binary judgment on the veracity of the news. Participants also reported their confidence in their judgement on a 7-point Likert scale from 1 (not confident at all) to 7 (extremely confident). Finally, the participant was asked to indicate how likely they would share this news through their social media accounts, again on a 7-point Likert scale between 1 (impossible to share) to 7 (extremely likely to share).

Figure 1 shows an example of our task interface. Note that participants received social influences by reviewing others' opinions about the news[2]—When the preceding participants who previously reviewed the news included both people with health-related domain expertise and those without, the current participant would be affected by both peer influence and expert influence (see details in Section 3.2.2). We acknowledge that in practice, the veracity of some news may not be either completely real or completely false but mixed [57, 102], especially when the news article is long. However, since the health-related news headlines used in this study make concise claims and can be objectively verified as either true or false, we opted to have subjects make binary veracity judgments on the news instead of more nuanced judgements (e.g., probabilistic estimates of veracity [37, 108]), and we left the investigation of the news with mixed truth to future studies.

3.2 Experimental Design

3.2.1 Experimental treatments. By varying the presence of the AI-based credibility indicator, we created two treatments:

- **Control (No AI)**: Participants in this treatment did *not* see the AI-based credibility indicator when reviewing news in each task.
- **Experimental (AI presented)**: In each task, along with the news and the preceding participants' opinions of its veracity, participants in this treatment would also see an AI model's prediction on the veracity of the news. Participants made their veracity judgement after reviewing all this information.

For participants assigned to treatments where AI is presented, we decided to display AI-based credibility indicators along with the news to mimic the action that many social media platforms (e.g., Twitter/X.com, Meta) have taken to signal false content, i.e.,

[1]The pre-registration documents for Experiment 1 can be found at https://aspredicted.org/779s-85s6.pdf. All experiments in this study are approved by the IRB of the authors' institution.

[2]In this experiment, we had participants explicitly expressed their opinions on the news veracity and showed these opinions to others. In practice, while people may not always indicate their precise judgement about the news veracity, they may make comments about a news story that indicate their belief in its veracity [12].

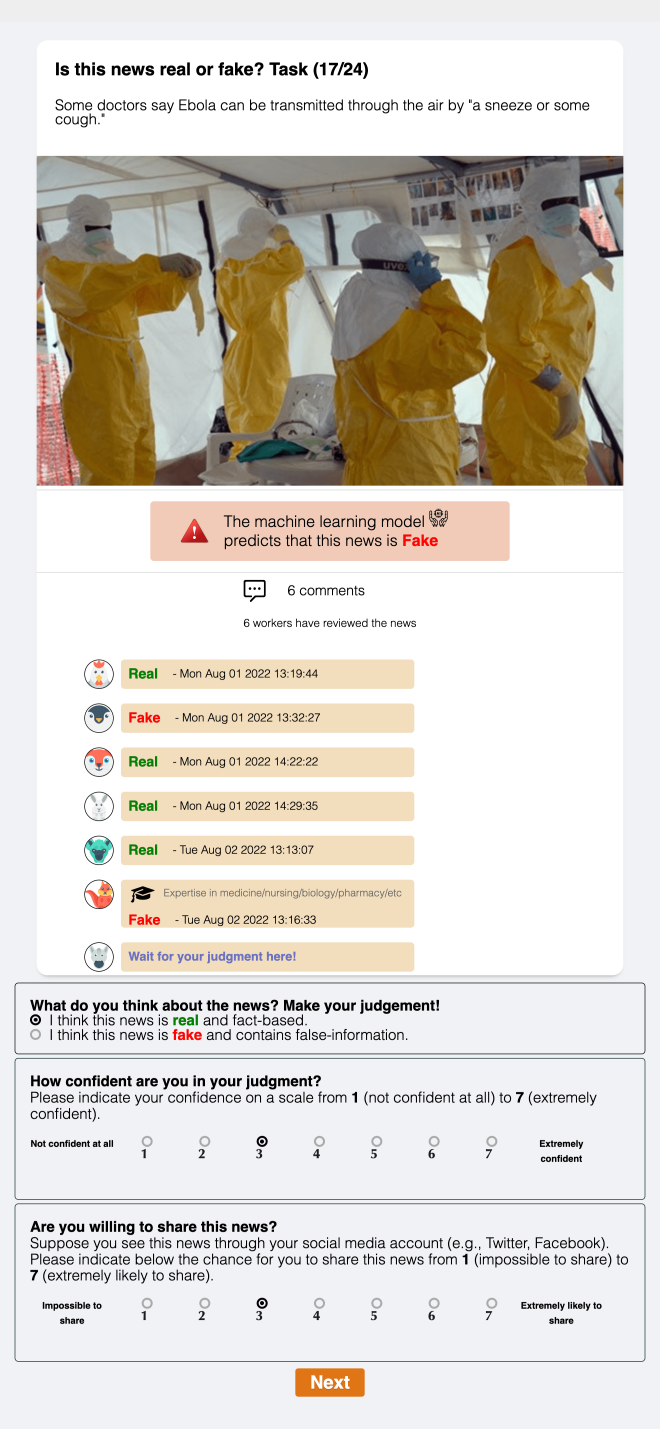


Figure 1: An example of the task interface. A graduation cap icon was placed along with the expert judgement to differentiate it from judgements made by peers.

directly displaying fact-checking warning labels along with the news to shape people's belief about the credibility of the news. In Experiment 1, the AI model we used was an oracle that *always provided the correct prediction* on the veracity of each news (i.e., its accuracy was 100%), although we did *not* inform participants about the accuracy of the model throughout this experiment. Since tasks like automatically assessing information credibility involve relatively high stakes, it is reasonable to expect that AI systems designed for these tasks need to achieve a high level of performance before they can be deployed in the real world. Recent research has also shown that state-of-the-art AI models can achieve an accuracy of 85+% in detecting misinformation [52, 62, 106, 133]. We note that by using a perfectly accurate AI model in Experiment 1, we can understand the highest level of benefits (i.e., the "upper bound") that can be brought up by AI-based credibility indicators in reducing misinformation when people are subject to influences from both peers and experts, and we will relax this assumption later in Experiment 3.

3.2.2 Incorporating the expert influence via a two-phase design. Recall that we aim to understand the effects of AI-based credibility indicators when people are influenced by both peers and *experts* (i.e., **RQ1**). This means that our experimental design needs to ensure that a significant portion of participants in our experiment should see veracity judgements made by "experts" in the task. Moreover, to enable an investigation into the effects of AI-based credibility indicators when they agree or disagree with the experts (i.e., **RQ2**), we also need to ensure the sample sizes are large enough for both scenarios. In light of this, in our experiments, instead of collecting real expert judgements, we adopted a two-phase design to incorporate *artificially-generated* "expert judgments" into the tasks (see Figure 2 for a schematic diagram of the two-phase design).

In particular, for each news, **Phase 1** was used to collect subsequent veracity judgements made by participants who reviewed this news, where each participant was only influenced by their peers (i.e., the preceding participants) and possibly the AI model, but *not* the experts when making their judgements; we referred to the sequence of veracity judgements generated by them as the "*pre-expert sequence*." In reality, a piece of news can get spread in the social network along multiple "paths." To simulate this, in Phase 1, we collected 4 pre-expert sequences for each of the 40 news in our dataset for each treatment, resulting in $40 \times 4 = 160$ pre-expert sequences per treatment. For example, given a Phase 1 participant who was assigned to the AI PRESENTED treatment, on each task, we would randomly select one of the 160 pre-expert sequences from the AI PRESENTED treatment (while ensuring that the participant reviewed different news in different tasks). The news and the preceding participants' veracity judgements recorded in the selected sequence would then be shown to the participant, as well as the AI model's prediction on the veracity of this news. The participant would then make their own veracity judgement, indicate their confidence, and express their willingness to share this news. Finally, this participant's veracity judgement would be appended to the end of the selected pre-expert sequence and be viewed by later participants who received this same sequence. By the end of Phase 1, we ensured that each pre-expert sequence contained at least 7 veracity judgements.

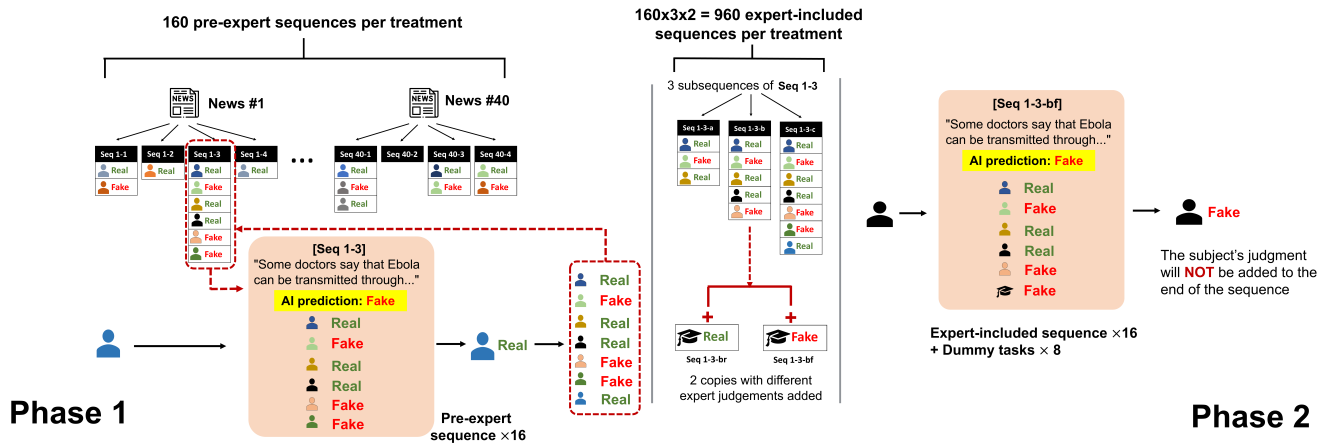


Figure 2: A schematic diagram of our two-phase experimental design. The AI-based credibility indicator (i.e., the part in yellow) will only be shown to participants who are assigned to the AI PRESENTED treatment.

Upon completion of Phase 1, we modified the pre-expert sequences to include artificial expert judgements. Specifically, for each pre-expert sequence we collected in Phase 1, we took three subsequences of it by preserving only the first 3, 5 or 7 judgements. Then, for each subsequence, we made two copies of it—for both copies, we appended a single "expert judgement" to the end of the subsequence, with the expert judgement agreeing with the AI model (hence correct) in one copy and disagreeing with the AI model (hence incorrect) in another copy[3]. That is, after inserting expert judgements, for each treatment, we had $40 \times 4 \times 3 \times 2 = 960$ sequences of veracity judgements with varying length (4, 6, or 8), and the last judgement in the sequence was always an expert judgement[4]. We call these sequences "*expert-included sequence*."

Finally, we used **Phase 2** to collect participants' veracity judgements and sharing intention on news, when they were subject to both peer influence and expert influence. Specifically, for a Phase 2 participant, we would randomly select one of the 960 expert-included sequences of the participant's assigned treatment on each task (while ensuring that the participant reviewed different news in different tasks). To help participants differentiate peer judgements (i.e., those made by laypeople participants in Phase 1) from the "expert judgement" (created by us artificially) on the news, we put a graduation cap icon along with the expert judgement as well as a note indicating that the person making the expert judgement has expertise in health-related domains like medicine, nursing, biology, and pharmacy (see Section 3.3 for details). Moreover, unlike that in Phase 1, Phase 2 participants' own veracity judgements would *not* be appended to the end of the expert-included sequence.

With this two-phase design, we can later focus on *only the Phase 2 participants* in the two treatments to examine whether and how the presence of AI-based credibility indicators affects people's perceptions of and engagement with online information when they are influenced by both peers and experts.

[3]This implies that expert judgements are independent with AI predictions; this is plausible as experts may be confident in themselves and less susceptible to influence.
[4]We varied the number of peer judgements before the expert judgement in our experiments to increase the generalizability of our results.

3.3 Experimental Procedure

We posted our Experiment 1 as human intelligence tasks (HITs) on Amazon Mechanical Turk (MTurk) to U.S. workers only, and we allowed each worker to take the HIT at most once. For participants in both phases of Experiment 1, upon arrival, they were first randomly assigned to one of the two treatments, and they were asked to pick an avatar to represent themselves throughout the experiment. Then, Phase 1 participants completed 16 tasks in the HIT, and in each task the news the participant reviewed was decided by the pre-expert sequence that we randomly drew from the participant's treatment (we also ensured that the participant saw different news in different tasks). On the other hand, Phase 2 participants completed 24 tasks in the HIT.

Importantly, for Phase 2 participants, we told them that if a graduation cap icon was presented next to a preceding participant's veracity judgement, it means that participant *claimed* to have health-related expertise themselves. To make this more credible, each participant in our experiment started the HIT by completing a demographics survey in which they were asked in one question if they have expertise in medicine/nursing/biology/pharmacy. However, we note that whether a graduation cap icon was shown along with a judgement was actually decided by whether that judgement was an expert judgement artificially generated by us instead of the participant's survey response, though participants were not aware of this.

Among the 24 tasks that a Phase 2 participant worked on, 16 tasks were *real tasks* in which the news the participant reviewed was decided by the expert-included sequence that we randomly selected from their treatment. When selecting these sequences, we also ensured that different tasks had different news, and the news veracity as well as the agreement between the expert judgement and the AI prediction were balanced across the 16 tasks (i.e., 4 cases each for the 4 scenarios: real news and the expert agrees with AI, real news and the expert disagrees with AI, fake news and the expert agrees with AI, fake news and the expert disagrees with AI). The fact that the real tasks always involved some judgement

from experts and each task had exactly one expert judgement may lead to participants' suspicions that they were working on a controlled experimental study, which may influence their behavior. To mitigate these suspicions, we included another 8 *dummy tasks* for participants to work on—We used a different set of 8 health-related news for these 8 dummy tasks[5]. For each dummy task, we randomly generated the preceding veracity judgements that the participant would see, among which the number of expert judgements to be shown (if any) and the positions of the expert judgements in the sequence were also randomly determined.

For participants in both phases, we provided a debrief to them revealing the veracity of each news they had reviewed after they completed all tasks in the HIT. Moreover, for participants in Phase 2, during the debrief, we also communicated to them that the expert judgements they saw in the tasks were actually generated by us artificially instead of being provided by other MTurk workers who self-claimed to have health-related expertise.

The base payment of our HIT was $1.6. To encourage our participants to carefully review the news and make accurate judgments, we also provided them with a performance-based bonus: We paid an extra 5 cents for each correct veracity judgment the participant made if their overall accuracy exceeded 65%. Thus, Phase 1 participants (Phase 2 participants) could receive a bonus of up to $0.8 ($1.2), in addition to the base payment. To help us later filter out inattentive participants, we also included an attention check question in the HIT in which participants were instructed to select a predefined option.

3.4 Analysis Methods

Independent and Dependent Variables. The main independent variable used in our analysis is the treatment assigned to participants, i.e., the presence of the AI-based credibility indicators.

To understand how the presence of the AI-based credibility indicators affects people's capability in detecting misinformation, we pre-registered two dependent variables: (1) the *accuracy* of a participant's judgment on the news veracity, and (2) a participant's *truth discernment*, which is calculated as the participant's frequency of labeling real news as "real" minus the their frequency of labeling fake news as "real." This metric is widely used in previous research [29, 87, 90] as it reflects people's sensitivity in differentiating real and fake news.

Similarly, to understand the effects of AI-based credibility indicators on the spread of misinformation, we pre-registered a few additional dependent variables: (1) a participant's self-reported *sharing intention* for the real or fake news, and (2) a participant's *sharing discernment*, which is calculated as the participant's sharing intention on real news minus that on fake news.

Intuitively, the presence of AI-based credibility indicators can help reduce misinformation if they can increase participants' veracity judgement accuracy, truth discernment, and sharing discernment, and nudge participants into being more willing to share real news and less willing to share fake news. Note that all dependent variables are measured using the experimental data collected from *Phase 2 participants* on the *real tasks* only.

[5]Through the pilot study, we again found that people's independent veracity judgement on these 8 news are between 50% and 60%, suggesting limited prior knowledge.

Statistical Methods. In Experiment 1, we focus on answering **RQ1** and **RQ2** for the scenario that an *ideal*, extremely accurate AI-based credibility indicator is used, while people's processing of online information is influenced by peers and *self-claimed* experts. Specifically, to answer **RQ1**, for each dependent variable that we have described above, we focus on the data obtained from Phase 2 participants and conduct t-tests between participants of the two treatments. Then, to answer **RQ2**, we split the data into two subgroups based on whether the Phase 2 participant observed the AI model's veracity judgement to be the same as the expert's judgement or not. For participants in the control (No AI) treatment, despite that they did not actually see the AI model's prediction, we still divided their data into two subgroups based on whether the expert judgement they saw in a task would be the same as the AI model's prediction should it be presented; this allowed us to compute the reference values of dependent variables for the control treatment when participants were only influenced by the expert and their peers. We then conducted t-tests between the two treatments within each subgroup of data. For both research questions **RQ1** and **RQ2**, we measured the effect sizes using Cohen's d.

3.5 Results

In total, 174 MTurk workers took our Phase 1 HIT and passed the attention check (control: 87, experimental: 87). Then, 201 workers took our Phase 2 HIT and passed the attention check (control: 92, experimental: 109). In the following, we analyze the data obtained from Phase 2 of the experiment to examine whether an ideal AI-based credibility indicator can help reduce misinformation when people are subject to both peer influence and self-claimed-expert influence.

RQ1: Effects on the Detection and Spread of Misinformation. We start by examining whether, overall, the presence of AI-based credibility indicators can help reduce misinformation when people are influenced by peers as well as experts with self-claimed domain expertise. First, we look into whether providing people with AI-based credibility indicators can help them detect misinformation more accurately. Figures 3a and 3b (the "All" row) compare participants' accuracy in judging news veracity and their truth discernment, respectively, across the two treatments. It is clear that when people are influenced by the opinions of other peers and self-claimed experts, the presence of accurate AI-based credibility indicators can still increase both their accuracy in identifying misinformation and their sensitivity in differentiating true and false information. Our t-test results further confirm that the increases are statistically significant in both the veracity judgement accuracy (Cohen's $d = 0.35$, $t(3215) = 10.00$, $p < 0.001$) and truth discernment (Cohen's $d = 1.04$, $t(200) = 7.22$, $p < 0.001$).

Next, we move on to examine whether the presence of AI-based credibility indicators has any impact on people's intention to spread the news when they are influenced by peers and experts with self-claimed domain expertise. The "All" row in Figures 3c and 3d show participants' self-reported willingness to share real news and fake news respectively. Compared to participants in the control treatment who did not receive the AI-based credibility indicators, participants in the AI PRESENTED treatment significantly increased their willingness to share real news (Cohen's $d = 0.12$,

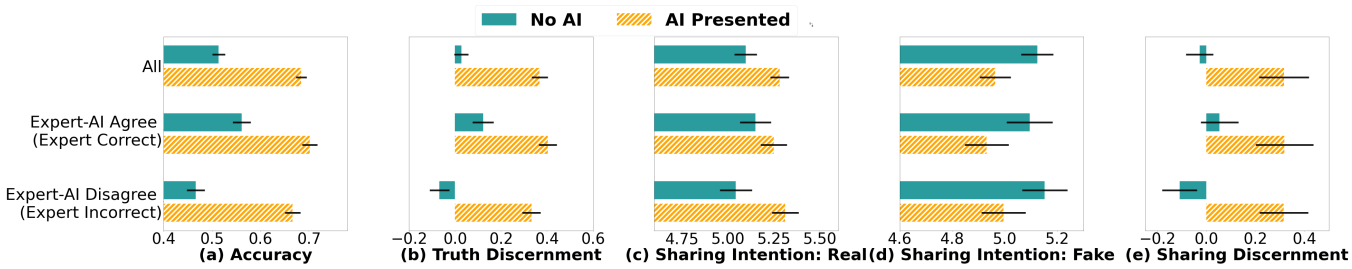


Figure 3: The impacts of AI-based credibility indicators on participants' ability in detecting misinformation and their intention to spread true and false information, when they are influenced by peers and experts with self-claimed expertise (Experiment 1). Results are presented for three conditions: (1) on all the data (All), (2) on a subset of the data where the experts' veracity judgement is the same as that of the AI model's (Expert-AI Agree), and (3) on a subset of the data where the experts' veracity judgement is different from the AI model's (Expert-AI Agree). Error bars represent the standard errors of the mean.

$t(1607) = 2.34, p = 0.002$). In addition, participants in the AI PRESENTED treatment also became less willing to share fake news, although our t-test result suggests that this decrease is marginal ($t(1607) = 1.87, p = 0.062$). Nevertheless, when we consider the extent to which people share more real news in relative to fake news (i.e., sharing discernment), as shown in Figure 3e, the presence of AI-based credibility indicators results in a significant increase in sharing discernment (Cohen's $d = 0.41$, $t(200) = 3.37, p < 0.001$).

Together, these results suggest that when people are influenced by both peers and self-claimed experts as they process online information, the presence of an extremely accurate AI-based credibility indicator can significantly improve people's capability in detecting misinformation and reduce their spreading of misinformation.

RQ2: Does Expert-AI Agreement Change the Effects of AI-based Credibility Indicators? We now move on to answer **RQ2**, i.e., examining whether the effects of AI-based credibility indicators are the same in the two scenarios where the AI model's veracity predictions of the news agree or disagree with the expert's judgements. In Figure 3, the "Expert-AI agree" and "Expert-AI disagree" rows compare the veracity judgement accuracy, truth discernment, willingness to share real and fake news, and sharing discernment between participants of the two treatments, for the "Expert-AI agree" and "Expert-AI disagree" scenarios separately.

We start by analyzing the scenario where the AI model's prediction is the same as the expert's judgement (i.e., the "Expert-AI Agree" bar in Figure 3). In terms of people's ability to detect misinformation (Figures 3a–3b), we find that when the AI model and the expert agree with each other, the presence of the AI-based credibility indicators help people further increase their veracity judgement accuracy (Cohen's $d = 0.29$, $t(1607) = 5.91, p < 0.001$) and enhance their truth discernment (Cohen' $d = 0.66$, $t(200) = 4.67, p < 0.001$). Recall that in this experiment, the AI model's predictions are always correct. Thus, these results effectively suggest that when people's ability in detecting misinformation is already positively influenced by some expert's *correct* judgements on news veracity, the explicit provision of an AI-based credibility indicator that *agrees* with the expert will bring about a significantly larger positive influence. In contrast, with respect to how the presence of AI-based credibility indicators affects people's willingness to engage with the news when the AI predictions are consistent with the expert judgements (Figures 3c–3e), we are only able to detect a marginal increase in sharing discernment ($p = 0.07$).

For the scenario where the AI model's prediction is different from the expert's judgement (i.e., the "Expert-AI Disagree" row of bars in Figure 3), we again detect significant increases in people's veracity judgement accuracy (Cohen's $d = 0.41$, $t(1607) = 8.26, p < 0.001$) and truth discernment (Cohen's $d = 0.97$, $t(200) = 6.83, p < 0.001$) due to the presence of the AI predictions. In addition, the presence of AI-based credibility indicators also result in a significant increase in people's willingness to share real news (Cohen's $d = 0.17, t(803) = 2.39, p = 0.017$) and their sharing discernment (Cohen's $d = 0.49, t(803) = 3.37, p < 0.001$). In other words, while people can be misled by some expert's *incorrect* judgements on news veracity, the explicit provision of an AI-based credibility indicator that *disagrees* with the expert can mitigate the negative expert influence while establishing a positive impact on both people's perceptions of and engagement with the news.

To formally compare the effect sizes between the two scenarios, we conducted bootstrap re-sampling ($K = 1000$) within each subgroup of data (i.e., the "Expert-AI agree" subgroup and the "Expert-AI disagree" subgroup). Given a bootstrapped sample of the data, we estimated the effect size of the impacts of AI-based credibility indicators using Cohen's d, for each of the three dependent variables for which at least marginal effects were detected in both scenarios (i.e., accuracy, truth discernment, and sharing discernment). We then used the paired t-test to compare the mean value of the estimated effect sizes in the Expert-AI agree scenario with that in the Expert-AI disagree scenario, and results are reported in Table 2. We find that when the expert's judgement disagrees with the AI model's prediction, the presence of the AI prediction consistently exhibits a larger impact on all three dependent variables. This implies that the provision of correct AI-based credibility indicators can be especially powerful in correcting the negative influence brought up by some expert's incorrect veracity judgement.

Put together, our Experiment 1 results suggest that the presence of an ideal, perfectly accurate AI-based credibility indicator can help people detect misinformation more accurately regardless of the agreement between experts and AI, but its impacts on people's

Dependent Var	d (Expert-AI Agree)	d (Expert-AI Disagree)	$\Delta\overline{d}$
Accuracy	0.28 [0.21, 0.34]	0.39 [0.32, 0.47]	-0.11***
Truth discernment	0.67 [0.44, 0.94]	0.97 [0.73, 1.22]	-0.30***
Sharing discernment	0.26 [0.06, 0.45]	0.49 [0.29, 0.68]	-0.24***

Table 2: 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 1. $\Delta\overline{d}$= d(Expert-AI agree) - d(Expert-AI disagree) is the difference of the average effect sizes. * represents a significance level of 0.001.**

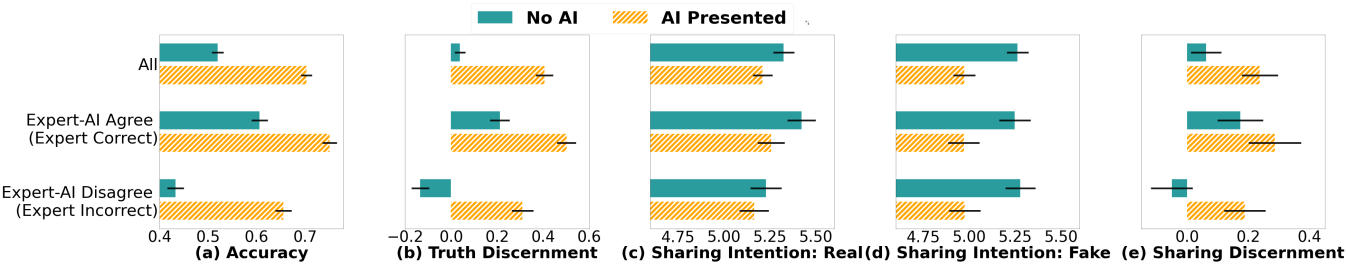


Figure 4: The impacts of AI-based credibility indicators on participants' ability in detecting misinformation and their intention to spread true and false information, when they are influenced by peers and experts with verified expertise (Experiment 2). Error bars represent the standard errors of the mean.

willingness to share the news are only reliable when the AI predictions disagree with the expert judgements (see the supplemental materials for additional analysis showing that the observed effects are robust to the variations in the number of peer judgements that a participant saw). Moreover, the impact of the AI-based credibility indicators on people is found to be always larger when the expert disagrees with the AI than when they agree with each other.

4 Experiment 2: When Expert's Expertise is Verified

Experiment 1 shows the effects of ideal AI-based credibility indicators on people's processing of online information when they are influenced by both peers and *self-claimed* experts. However, in reality, many social media platforms (e.g., Twitter/X.com, Sina Weibo, and some subreddits in Reddit) provide verification to expert users and present their accounts in a different way than other users (e.g., Twitter verification icon). Arguably, when the experts are in some ways "verified" by the platforms, their influence to other people may become larger compared to when they simply claim that they are experts in some domain. As a result, in Experiment 2 [6], we conduct another study to explore the generalizability of our findings in Experiment 1, when people's perception of and engagement with information is under the mixed influence of peers and experts with *verified* expertise.

4.1 Experimental Design and Procedure

We adopted the same design and procedure of Experiment 1 in our Experiment 2, except for making a few minor changes: (1) We reused the pre-expert sequences obtained from Phase 1 of Experiment 1 and only collected Phase 2 data for Experiment 2; (2) We told participants that in addition to collecting veracity judgements from MTurk workers, we also recruited researchers (e.g., postdoctoral researchers, senior graduate students, etc.) to review the news, and we verified that these researchers hold degrees in health-related disciplines like medicine, nursing, biology, and pharmacy; whenever a veracity judgement was displayed along with a graduation cap icon and a verified check icon, it was made by one of these researchers; (3) Participants of Experiment 1 were excluded from taking part in this experiment.

4.2 Results

In total, we obtained Phase 2 data from 208 valid workers in Experiment 2 (control: 106, experimental: 102). In the following, we revisit RQ1 and RQ2 with the Phase 2 data of Experiment 2 to answer **RQ3**, i.e., to understand whether AI-based credibility indicators can still help reduce misinformation when people are subject to influence from both peers and *verified* experts.

RQ3: Effects of AI-based Credibility Indicators When Experts Are Verified. Figure 4 compares people's ability in detecting misinformation and their intention to engage with true and false information across subjects in the two treatments of Experiment 2.

First, we focus on understanding the overall effects of the AI-based credibility indicators regardless of the agreement between experts and AI (i.e., the "All" row in Figure 4). We still find the presence of the AI-based credibility indicator significantly increases people's capability in detecting misinformation even as they are influenced by peers and experts with verified expertise (accuracy: Cohen' $d = 0.38$, $t(3327) = 11.07, p < 0.001$; truth discernment: Cohen's $d = 1.15$, $t(207) = 8.30, p < 0.001$). In addition, while

[6]The pre-registration documents for Experiment 2 can be found at https://aspredicted.org/3f79-n5bp.pdf

the presence of AI-based credibility indicators show no impact on people's willingness to share real news, it significantly decreases people's willingness to share fake news (Cohen's $d = 0.17$, $t(1663) = 3.43, p < 0.001$) and increases people's sharing discernment (Cohen's $d = 0.32$, $t(207) = 2.29, p = 0.023$).

These effects of AI-based credibility indicators largely hold true when we take a closer look into the two scenarios where the AI model's predictions agree or disagree with the experts' judgements separately. For example, as shown in Figures 4a–4b, with the presence of AI-based credibility indicators, participants' accuracy in judging news veracity and their truth discernment are significantly increased regardless of the expert-AI agreement (accuracy: *Expert-AI Agree*: $p < 0.001$, Cohen's $d = 0.31$, *Expert-AI Disagree*: $p < 0.001$, Cohen's $d = 0.46$; truth discernment: *Expert-AI Agree*: $p < 0.001$, Cohen's $d = 0.68$, *Expert-AI Disagree*: $p < 0.001$, Cohen's $d = 1.01$). Moreover, in both scenarios, we do not find significant impacts of the AI-based credibility indicators on people's willingness to share real news (Figure 4c), but we do detect significant impacts on people's willingness to share fake news (Figure 4d; *Expert-AI Agree*: $p = 0.02$, Cohen's $d = 0.16$; *Expert-AI Disagree*: $p = 0.01$, Cohen's $d = 0.18$). In terms of people's sharing discernment (Figure 4e), we find that the presence of AI-based credibility indicators only make people share more real news in relative to fake news when the verified experts' judgements disagree with the AI predictions ($p = 0.01$, Cohen's $d = 0.35$). Finally, via comparing the sizes of the AI predictions' effects, we again conclude that the presence of AI-based credibility indicators exerts a larger impact on people's detection and spread of misinformation when the verified expert *disagrees* with AI (see the supplementary material for more details).

Together, results from Experiment 2 confirmed the effectiveness of an ideal AI-based credibility indicator in helping people detect misinformation and preventing people from sharing misinformation, even when both laypeople peers and verified experts influence them. Again, these effects are consistently larger when the judgements of the verified experts disagree with the AI prediction.

5 Experiment 3: When AI Accuracy Varies

Our Experiments 1 and 2 have thoroughly investigated how ideal AI-based credibility indicators could help people better detect misinformation and appropriately engage with information, when people are under the social influence of a mixed crowd of laypeople and experts. Despite the promise of utilizing advanced AI technologies to combat misinformation, AI models may still make mistakes in practice. Therefore, in our final Experiment 3 [7], we aim to understand how the impacts of AI-based credibility indicators on people's perception and engagement of online information change, as the accuracy of the AI model underlying the credibility indicator changes.

5.1 Experimental Design and Procedure

We largely followed the design and procedure of Experiment 1 in Experiment 3. The main difference is that in Experiment 3, in order to take the AI-based credibility indicator's varying accuracy into account, we created the following three treatments:

- **Control (No AI)**: Participants in this treatment did *not* see the AI-based credibility indicator when reviewing news in each task.
- **High Accuracy AI**: In each task, along with the news and the preceding participants' opinions of its veracity, participants in this treatment would also see an AI model's prediction on the veracity of the news. The accuracy of the AI model is 80%.
- **Low Accuracy AI**: Same as that in the previous treatment, an AI model's veracity prediction is presented along with the news and preceding participants' opinions. However, the accuracy of the AI model in this treatment is 55%.

The AI model used in the High Accuracy AI treatment is implemented via the OpenAI LLM API empowered by GPT-3.5 turbo, using a simple prompt instructing the language model to classify whether the news is fact-based (real news) or contains false information (fake news). On the other hand, in the Low Accuracy AI treatment, we trained a multinomial Naive Bayesian classifier on a health-related news dataset and utilized it as the AI model. Note that we set the accuracy of AI model at 80% in our High Accuracy AI treatment to understand the effects of a decently accurate AI-based credibility indicator which still makes a considerable number of incorrect judgements.

Similar to that in Experiment 1, we told participants that the expert judgements they saw were made by people who self-claimed to have health-related expertise themselves. In addition, participants of both Experiments 1 and 2 were excluded from taking part in Experiment 3.

5.2 Results

In total, 644 MTurk workers attended our Experiment 3, with 271 workers (control: 87, High-accuracy AI: 90, Low-accuracy AI: 94) taking the Phase 1 HIT and 373 workers (control: 127, High-accuracy AI: 123, Low-accuracy AI: 123) taking the Phase 2 HIT, respectively. We then analyze the Phase 2 data of Experiment 3 to answer **RQ4**, i.e., how the AI-based credibility indicator's impact on people's perception of and engagement with misinformation varies with the AI model's accuracy, when people are influenced by both peers and self-claimed experts.

RQ4: Effects of AI-based Credibility Indicators When AI Accuracy Varies. First, we look into whether providing imperfect AI-based credibility indicators can still help people detect misinformation more accurately than they did independently. Figures 5a and 5b compare participants' accuracy in judging news veracity and their truth discernment, respectively, across the three treatments. Our ANOVA results suggest that there is a significant difference across treatments on both participants' veracity judgement accuracy and truth discernment (accuracy: $F(2, 5965) = 11.2, p < 0.001$, truth discernment: $F(2, 370) = 11.62, p < 0.001$). Furthermore, the post-hoc Tukey HSD pairwise comparisons suggest that participants in the High Accuracy AI treatment are significantly more accurate in evaluating the veracity of different news and in discerning true and false information than both those participants who did not have access to the AI-based credibility indicators (accuracy: $p = 0.007$, Cohen's $d = 0.10$; truth discernment: $p = 0.007$, Cohen's $d = 0.37$) and those in the Low Accuracy AI treatment (accuracy:

[7] The pre-registration documents for Experiment 3 can be found at https://aspredicted.org/sfz5-7rf7.pdf

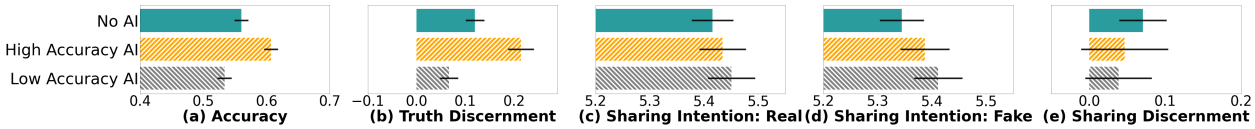


Figure 5: The impacts of AI-based credibility indicators of different accuracy on participants' ability in detecting misinformation and their intention to spread true and false information, when they are influenced by peers and experts with self-claimed expertise (Experiment 3). Error bars represent the standard errors of the mean.

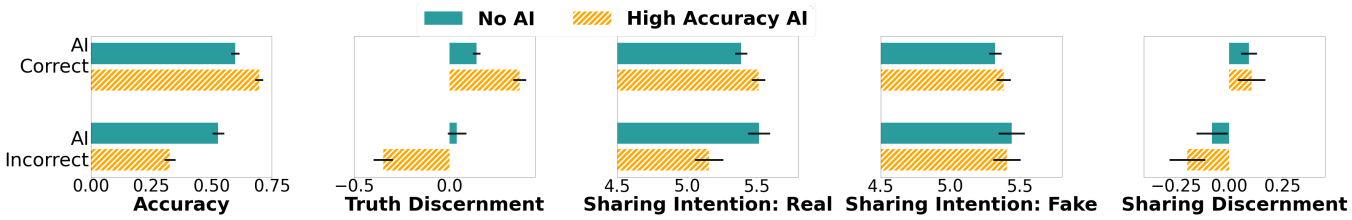


Figure 6: The impacts of high accuracy AI-based credibility indicators on participants' ability in detecting misinformation and their intention to spread true and false information, when they are influenced by peers and experts with self-claimed expertise (Experiment 3). Data is separated into two subgroups based on whether the prediction of the AI model used in the High Accuracy AI treatment is correct or incorrect. Error bars represent the standard errors of the mean.

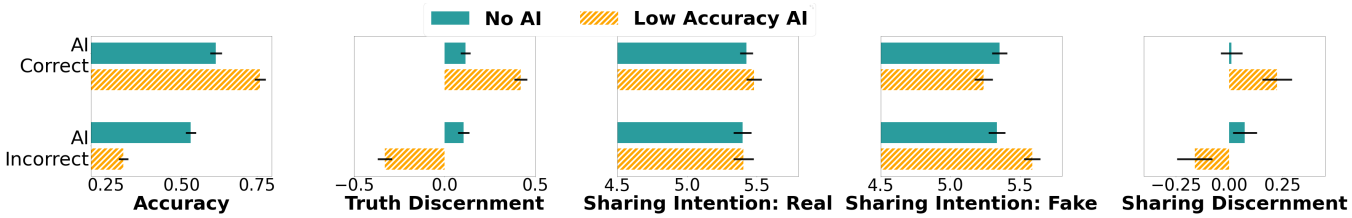


Figure 7: The impacts of low accuracy AI-based credibility indicators on participants' ability in detecting misinformation and their intention to spread true and false information, when they are influenced by peers and experts with self-claimed expertise (Experiment 3). Data is separated into two subgroups based on whether the prediction of the AI model used in the Low Accuracy AI treatment is correct or incorrect. Error bars represent the standard errors of the mean.

$p < 0.001$, Cohen's $d = 0.15$; truth discernment: $p < 0.001$, Cohen's $d = 0.58$).

Moreover, Figure 5c–5e compare participants' self-reported willingness to share real news and fake news, and their sharing discernment across the three treatments of Experiment 3. Interestingly, we did not see clear patterns in terms of the effects of the AI-based credibility indicator regardless of its accuracy. Our one-way ANOVA tests confirm that the presence of imperfect AI-based credibility indicator with varying levels of accuracy does not significantly change people's intention to share real news, fake news, or their sharing discernment.

To gain a deeper understanding of why we obtain these findings, we then move on to analyze the effects of correct and incorrect AI predictions separately. As high accuracy AI and low accuracy AI may make different predictions on the same news, it is difficult for us to conduct this analysis on the data of all three treatments together. As a result, we conduct this analysis between the control treatment and each of the two experimental treatments separately (i.e., High Accuracy AI vs. Control, Low Accuracy AI vs. Control).

Figure 6 (the "AI correct" row) shows the effects of the AI-based credibility indicator with high accuracy when AI makes correct predictions[8]. We find that when the AI prediction is correct, it improves people's capability in detecting misinformation and differentiating between true and false information (accuracy: Cohen's $d = 0.23, t(4732) = 6.58, p < 0.001$; truth discernment: Cohen's $d = 0.73, t(372) = 5.76, p < 0.001$). However, we did not find statistical evidence supporting that correct predictions of high accuracy AI could further help people engage more appropriately with information ($p > 0.05$ for sharing intention on both real and fake news, and sharing discernment). Oppositely, examining the "AI Incorrect" row in Figure 6, we found that the incorrect AI predictions severely mislead people to make incorrect judgements on news stories (accuracy: Cohen's $d = 0.41, t(1234) = 5.94, p < 0.001$), and become less capable of differentiating real and false information (truth discernment: Cohen's $d = 0.76, t(308) = 5.43, p < 0.001$).

[8]Participants in the control treatment did not see the AI model's prediction. Dependent variable values for the control treatment shown in Figure 6 (or Figure 7) were calculated for the two subsets of tasks where the AI model used in the "High Accuracy AI" (or "Low Accuracy AI") treatment made correct or incorrect predictions separately.

Experiment/Independent Var	y = Final Accuracy		
	Experiment 1	Experiment 2	Experiment 3
Intercept (C)	-0.08	-0.22^{***}	0.08
Expert Accuracy (β_2)	0.27^{***}	0.60^{***}	0.31^{***}
Correct AI Presence (β_3)	0.72^{***}	0.80^{***}	0.60^{***}
Incorrect AI Presence (β_4)			-1.02^{***}

Table 3: Regressions for understanding whether AI affects people's detection of and engagement with misinformation more than experts. *** represents a significance level of 0.001.

Moreover, when high accuracy AI misclassified real news as fake, it significantly reduced people's intention to share the news (Cohen's $d = 0.27, t(640) = 2.8, p = 0.005$).

Similarly, when we look into the case where low accuracy AI makes correct predictions (the "AI Correct" row in Figure 7) and those where low accuracy AI makes incorrect predictions (the "AI Incorrect" row in Figure 7), we obtained similar findings: correct AI predictions enhance people's capability in detecting misinformation (Cohen's $d = 0.29, t(3272) = 6.77, p < 0.001$) and differentiating real and fake news (Cohen's $d = 0.86, t(371) = 6.78, p < 0.001$), while incorrect AI predictions led to incorrect judgements (accuracy: Cohen's $d = 0.46, t(2694) = 9.83, p < 0.001$; truth discernment: Cohen's $d = 1.09, t(372) = 8.57, p < 0.001$). Furthermore, correct predictions of low accuracy AI can enhance people's intention to engage more with real information than false information (Cohen's $d = 0.86, t(371) = 6.78, p < 0.001$), while incorrect predictions can undermine it (Cohen's $d = 0.30, t(372) = 2.36, p = 0.019$).

Together, these results suggest that when people are influenced by both peers and self-claimed experts in processing online information, imperfect AI-based credibility indicators can still boost people's capabilities in recognizing and differentiating real and fake news if the AI model is relatively highly competent. However, when the accuracy of the AI is very low, not only will it make no positive impact but it also brings risks to people's capability in judging news veracity. On the other hand, under the mixed influence from other people and self-claimed experts, the imperfect AI-based credibility indicators does not appear to influence people's intention to share real or fake news anymore. By taking a deeper look into the effects of AI-based credibility indicators on cases where AI makes correct or incorrect predictions separately, we found that people lack the capability to differentiate correct and incorrect AI predictions, regardless of the level of performance of the AI. As a result, correct AI indications lead to the enhancement of people's capability in assessing information credibility, while incorrect AI indications undermine such capability, and may together lead to a null effect.

6 Exploratory Analysis

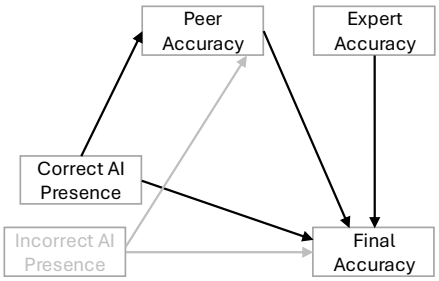
So far, our findings across the three experiments suggest that even when people are influenced by the opinions of both peers and experts in interpreting the news, providing a reasonably accurate AI-based credibility indicator along with the news can still enhance people's capability in detecting misinformation and even reduce people's engagement with misinformation. This holds true regardless of whether the expert is self-claimed or verified. However, it is challenging for people to distinguish correct AI predictions from the incorrect ones. To obtain deeper insights into the mechanisms underlying the impacts of AI-based credibility indicators on people, we conduct a few exploratory analyses.

6.1 Does AI Affect People's Detection of and Engagement with Misinformation More Than Experts?

First, we aim to understand how people weigh the opinions of different parties in their judgement of news veracity. People's judgement of news veracity can be influenced by the AI-based credibility indicator, the expert's opinion, as well as the opinions of the laypeople peers. Since laypeople peers' opinions can also be influenced by the AI model's predictions when they are presented, in this analysis, we focus on comparing the magnitude of the impacts of the two independent sources of influences—AI-based credibility indicators and experts.

To do so, we used logistic regression models to predict the accuracy of a participant's final veracity judgement in a task based on whether the expert's judgement that the participant saw was correct, and whether the participant had access to the AI-based credibility indicator in the task. In particular, for Experiments 1 and 2, we used a single independent variable "*Correct AI Presence*" to reflect the presence of the AI-based credibility indicator, as the indicator always provides correct predictions in these two experiments. In contrast, for Experiment 3, we used two independent variables "*Correct AI Presence*" and "*Incorrect AI Presence*" to indicate whether the participant received a correct AI prediction or an incorrect AI prediction in the task.

Regression results are shown in Table 3. We first note that across the three experiments, the presence of AI-based credibility indicator and the correctness of the expert's judgement both significantly influence the participant's veracity judgement accuracy ($p < 0.001$). Moreover, we notice that in Experiments 1 and 2, the increase in the participant's veracity judgement accuracy resulting from a correct expert judgement is consistently *smaller* than the increase brought up by the presence of the correct AI-based credibility indicator (i.e., $\beta_2 < \beta_3$). Similarly, in Experiment 3, the absolute magnitude of change in the participant's veracity judgement accuracy caused by the presence of a correct expert judgement is again *smaller* than that caused by either the correct or the incorrect AI prediction (i.e., $|\beta_2| < |\beta_3|$, $|\beta_2| < |\beta_4|$). This indicates that at least under our experimental setting, people's veracity judgements are influenced by AI to a larger degree. That said, we note that as the expert's expertise gets verified, their impacts become larger (i.e., β_2 is larger



Dependent Var / Independent Var	Model 1: y = Peer Accuracy			Model 2: y = Final Accuracy		
	Exp 1	Exp 2	Exp 3	Exp 1	Exp 2	Exp 3
Intercept (C)	-0.08	0.076	0.52^{***}	-0.88^{***}	-0.82^{***}	-0.87^{***}
Peer Accuracy (β_1)				1.60^{***}	1.17^{***}	1.88^{***}
Expert Accuracy (β_2)				0.37^{***}	0.61^{***}	0.36^{***}
Correct AI Presence (β_3)	0.71^{***}	0.65^{***}	0.19^{***}	0.53^{***}	0.67^{***}	0.31^{***}
Incorrect AI Presence (β_4)			-0.21^{***}			-0.80^{***}

Figure 8: Regressions for understanding whether the AI's impact on people's veracity judgement accuracy is mediated by the peers' veracity judgement accuracy. *** represents a significance level of 0.001. The variable "Incorrect AI presence" is only relevant for Experiment 3, and thus it is shown in grey in the hypothesized model.

in Experiment 2 than in Experiments 1 and 3) and closer to the impacts of AI.

Finally, we conduct similar regression analyses to compare the magnitude of the impacts of AI and experts in influencing participants' willingness to share a piece of news, and we again find that the impacts of AI-based credibility indicators are larger than those of the experts' (see the supplementary material for more details).

6.2 How does AI-based Credibility Indicators Affect People's Detection of and Engagement with Misinformation?

Next, we take a closer look into *how* AI-based credibility indicators affect people's veracity judgements—do they change a participant's veracity judgement by *directly* influencing the participant, or by *indirectly* influencing the veracity judgements of the laypeople peers who have reviewed the news before the participant (see left panel in Table 8 for a hypothesized model for the case that AI's impacts on people are mediated through peers)? To find out, we performed the mediation analyzes and the results are reported in Table 8 (right panel).

In particular, for each participant in our experiment, we first conducted regression analyses to check if the accuracy of the majority veracity judgements made by those preceding laypeople peers was influenced by the presence of AI-based credibility indicator. Similar as before, for Experiments 1 and 2, we used a single independent variable "*Correct AI Presence*" to reflect the presence of the correct AI-based credibility indicator, while for Experiment 3, we used two independent variables "*Correct AI Presence*" and "*Incorrect AI Presence*" to indicate whether a correct AI prediction or an incorrect AI prediction was presented to the participant as well as their preceding peers. Results of Model 1 in Table 8 suggest that in all three experiments, the presence of AI predictions significantly impacts the accuracy of the proceeding laypeople peers' veracity judgements ($p < 0.001$)—the presence of correct AI prediction significantly increases the peers' accuracy, while the presence of incorrect AI prediction significantly decreases the peers' accuracy.

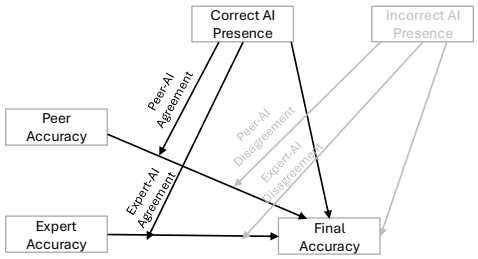
Moreover, in Model 2, we took the influences of all three parties (i.e., the laypeople peers, the expert, and the AI prediction) into consideration, to predict the participant's veracity judgement accuracy. As shown in Table 8, we find that the coefficients associated with both peer accuracy (i.e., β_1) and AI presence (i.e., β_3, β_4) are significant ($p < 0.001$). This means that the effects of AI-based credibility indicators on people's accuracy in judging news veracity are *partially* mediated by the peers' accuracy, i.e., the AI predictions impact people's ability to detect misinformation both directly and indirectly through peer influence (via changing peers' veracity judgements).

Finally, we also conduct similar mediation analyses to examine how the presence of AI-based credibility indicators affects people's willingness to share news, and we again find these impacts are partially mediated through peer influence (see the supplementary material for more details).

6.3 Will the Agreement between AI and Peers/Experts Change their Impacts on People?

When the AI-based credibility indicator is presented, it may agree or disagree with the majority opinions expressed by the laypeople peers. It may also agree or disagree with the expert's opinions. As such, a natural question to ask is whether the agreement or disagreement between AI and peers (or the expert) moderates the peers' (or the expert's) impacts on the participant's detection of misinformation. In other words, if correct peer (or expert) judgements in news veracity can increase the participant's accuracy in detecting misinformation, after a correct AI-based credibility indicator is presented and therefore the participant observed an agreement between AI and the peers (or the expert), will the magnitude of this increase change? What about in the case where an incorrect AI-based credibility indicator is presented and therefore the participant observed a disagreement between AI and the peers (or the expert)?

Table 9 (left panel) shows the hypothesized model for the case where the agreement between AI and peers/experts indeed moderates the impacts of peers/experts on people. To examine if this is the case, utilizing the experimental data collected from the three



Dependent Var / Independent Var	y = Final Accuracy		
	Exp 1	**Exp 2**	**Exp 3**
Intercept (C)	-1.14^{***}	-0.97^{***}	-1.03^{***}
Correct AI Presence (β_1)	1.05^{***}	1.00^{***}	0.68^{***}
Incorrect AI Presence(β_2)			-0.77^{**}
Peer Accuracy (β_3)	1.99^{***}	1.31^{***}	2.16^{***}
Expert Accuracy (β_4)	0.54^{***}	0.76^{***}	0.42^{***}
AI-Peer Agreement * Peer Accuracy (β_5)	-0.74^{***}	-0.32^{*}	-0.53^{**}
AI-Peer Disagreement * Peer Accuracy (β_6)			-0.26
AI-Expert Agreement * Expert Accuracy (β_7)	-0.31	-0.32^{*}	-0.19
AI-Expert Disagreement * Expert Accuracy (β_8)			0.16

Figure 9: Regressions for understanding how the agreement between AI and peers/experts moderates their impacts on people's veracity judgement accuracy. *,**,*** represent significance levels of 0.05, 0.01 and 0.001. The variable "Incorrect AI presence" is only relevant for Experiment 3, and thus it is shown in grey in the hypothesized model.

experiments, we used logistic regression models to predict a participant's accuracy in their news veracity judgement in a task based on two sets of variables: (1) the direct impacts of the presence of correct/incorrect AI predictions, whether the majority veracity judgement made by the preceding laypeople peers was correct, and whether the expert's veracity judgement was correct (i.e., terms associated with β_1–β_4); and (2) the potential moderating effects caused by the agreement between AI and peers/experts, including the interaction terms between peers' accuracy and the agreement/disagreement between AI and the peers (β_5, β_6), and the interaction terms between the expert's accuracy and the agreement/disagreement between AI and the expert (β_7, β_8)[9].

As results in Table 9 show, in all three experiments, we find a significantly negative interaction between AI-Peer agreement and the peers' accuracy (i.e., $\beta_5 < 0$ and is significant). This means that an agreement between the AI-based credibility indicator and the peers' opinions results in a *decrease* in the positive impacts of laypeople peers' correct veracity judgement on the participant's detection of misinformation. One possible explanation is that the participant knew that peers' veracity judgement may also be affected by the AI model, thus they automatically discounted the peer influences when seeing an agreement between the peers and AI. On the other hand, we also note that a disagreement between peers and AI does not significantly change the positive impacts of peers' correct veracity judgement on the participant's accuracy in detecting misinformation. This suggests that when the majority of the laypeople peers disagree with the AI prediction, participants may consider that the peers' judgements are of independent values. Additionally, the agreement between AI-based credibility indicator and the expert generally does not change the impacts of the expert, except when the expert's expertise is verified ($\beta_7 < 0$ and is significant in Experiment 2). This is possibly because when AI-based credibility indicators are not presented, participants might place high trust in the verified experts (as supported by the large estimate of β_4 in Experiment 2 compared to that in the other two experiments). However, as the AI-based credibility indicator became available and participants observed divergent opinions between the expert and AI, participants could significantly lowered their trust in the verified expert in general, which might have resulted in a spillover effect of decreased influences of the verified expert even when they agreed with AI.

[9]Note that "AI-Peer/Expert Agreement" is only set to 1 if the AI-based credibility indicator was presented in a task, and its prediction is the same as the opinions expressed by the majority of laypeople peers/the expert. Similarly, "AI-Peer/Expert Disgreement" is only set to 1 if the AI-based credibility indicator was presented in a task, and its prediction is different from the opinions expressed by the majority of laypeople peers/the expert. As a result, for Experiments 1 and 2, the regressions do not include interaction terms concerning "AI-Peer/Expert Disgreement" as they all equal to zero (in Experiments 1 and 2, when AI disagrees with peers/expert, peers/expert must be incorrect).

7 Discussion

In this section, we begin with outlining the potential benefits and risks of AI-based credibility indicators. Following that, we will explore the implications for better leveraging AI to counter misinformation. Finally, we discuss the limitations of our study.

7.1 Combating Misinformation with AI-based Credibility Indicator: Pros and Cons

The results of our study indicate that when individuals are subject to social influence from both laypeople peers and experts while judging the veracity of online information, leveraging AI-based credibility indicators offers several benefits, but also poses certain risks and limitations.

On the positive side, our research demonstrates that, despite the complexities of social influence consisting of laypeople peers and experts, the inclusion of accurate AI-based credibility indicators alongside news content can effectively enhance individuals' capacity to detect misinformation, no matter whether the expert is self-claimed or verified. This efficacy of AI-based credibility indicators is important, because the presence of expertise displayed on social media platforms can sometimes lead to a Halo effect, wherein individuals may overestimate the credibility of specific users who appear more persuasive than others. Yet, their judgments may be as uninformed as those of laypeople due to potential mismatches between their displayed expertise and the actual expertise required to assess the accuracy of news in specific domains. In some cases, these "experts" may even transition into the role

of social media influencers (SMIS), leading people to believe they possess specialized knowledge over the long term [33]. However, there is no guarantee that such SMIS will consistently provide accurate and verified information to their audience. As seen during the pandemic, some individuals with perceived knowledge or expertise propagated COVID-related misinformation or advocated for unreasonable actions against fact-based information [128]. Consequently, our results highlight the robustness of accurate AI-based credibility indicators in nudging most of laypeople peers to have a consistent judgment on news veracity. The fact that the effects of AI-based credibility indicators are larger when experts disagree with AI suggests the potential for people to avoid the Halo effect and calibrate their reliance on online experts through introducing a second opinion from AI.

However, AI-based credibility indicators come with inherent risks. First, individuals often struggle to discern the accuracy of AI predictions. Consequently, they can be easily swayed by incorrect AI assessments, which can impact their ability to appropriately engage with both true and false information. There also exists a potential risk that the disagreement caused by AI's incorrect predictions and real experts' correct judgements lead to people's decreased trust in the true experts, as people generally perform poorly in differentiating the correctness of AI predictions. Second, the fact that the influence of AI-based credibility indicators is mediated through the crowd presents additional risks when moving beyond individual's interactions with AI predictions. For instance, bots generating content tailored to these scenarios can create artificial social influence to either amplify the negative impacts of wrong AI predictions or minimize the positive impacts of correct AI predictions.

7.2 Implications for Designing Better AI-based Credibility Indicators to Combat Misinformation

In light of the significance of social influence in fully releasing the potential of AI-based credibility indicators in influencing people's perception of and engagement with online information, one may consider ways to further enhance the effectiveness of accurate AI-based credibility indicators by attaching "social proof" to these indicators. Just as metadata summarizing user interactions with social media posts has often been used to highlight the popularity of a particular post, metadata capturing people's agreement or disagreement with AI-based credibility indicators can also be utilized to amplify the influence of these indicators (e.g., by explicitly displaying the number/proportion of individuals who agree with AI on the interface). Another strategy involves recognizing the value of discrepancies between different information credibility sources, even for a mixture of reliable and unreliable sources such as AI predictions and expert judgments in our study. For example, Bayesian inference may be used to aggregate multiple credibility indicators together in an optimal way, considering the reliability for each one of them. Future studies should also look into how to best present multiple credibility indicators from various sources to people to inspire them to engage in more analytical thinking, rather than simply adopting a heuristic way to process these indicators and blindly follow a particular one.

Furthermore, the revealed risks of imperfect AI-based credibility indicators highlight the importance of requiring some guaranteed level of performance from AI models underlying the credibility indicators before deployment. Our study suggests that the AI-based credibility indicators tend to surpass even platform-verified experts in influencing people's perceptions of and engagement with online information. As a result, once AI makes a wrong prediction on a piece of news, it's challenging to rectify the perception of subsequent users of that news merely by relying on spontaneous remedies within the online community, such as the intervention of users with expertise providing factually correct judgments. This is because a stronger social influence led by AI might have already taken shape. This suggests that the issue of AI mistakes, as well as people's blind reliance on AI, may be a serious threat to the health of the online information environment. Beyond the necessity to build highly competent AI-based credibility indicators, it would also be helpful to have comprehensive education for users to appropriately utilize the AI-based credibility indicators and to establish a mechanism for the social media platform to respond to AI mistakes. Future work could also explore how to properly present the uncertainty quantification of AI-based credibility indicators to users, or to adaptively determine whether to present AI-based credibility indicators while accounting for the "implied truth effect" [85], in order to promote users' appropriate reliance on these indicators.

7.3 Limitations and Future Work

Our study has several limitations. To begin with, the experiments were conducted with crowdworkers (i.e., subjects recruited from Amazon Mechanical Turk) on one specific type of news (i.e., concise health-related news headlines), and the veracity of the news is binary. Cautions should be used when generalizing results in this work to news on different topics and among individuals with different characteristics. For example, the health-related news headlines we choose in this study have a clear binary ground truth regarding their veracity, and their content is generally not controversial (e.g., unlike some Covid-19 related news). In reality, the veracity of news can be mixed. An interesting future work is to explore that in these scenarios, how should the AI-based credibility indicators be properly designed (e.g., should the indicator simply indicate the veracity of the news is mixed, or specify how much or which part of the information is true?), and whether the presence of these indicators can still help people detect the veracity of the news. When news topics are more controversial, one may believe that the judgement on the veracity of news is "subjective" and that they may have a stronger emotional attachment to their veracity belief. Developing and appropriately evaluating an "accurate" AI-based credibility indicator for these news topics can be a significant challenge to be addressed on its own, and it is unclear how such an indicator would impact people's consumption of information on contentious topics. In addition, in a real-world social media environment, individuals are more likely to be connected with others who share similarities with themselves [69], and this may also influence what kind of information they tend to consume. For example, politics-related news with different political leaning tends to spread within different communities of users due to the echo chamber formed on the social media platforms [93, 110]. More experimental studies

should be conducted on a larger range of news topics where people hold stronger prior beliefs, and in more polarized settings where fewer disagreeing opinions arise among connected individuals. It is also interesting to explore whether the findings of this study hold in other domains where experts are generally perceived as significantly more or less credible than those in health.

Moreover, in reality, it is common for people to be influenced more by people closely connected to them; also, sometimes experts are, at the same time, social influencers or authorities so that they have larger impacts on their followers. In our experiments, judgments made by other crowdworkers and artificial experts may not sufficiently reflect different individuals' connectivity and reputations on social media. Therefore, it is unclear to what extent the conclusion can be generalized to the information diffusion process where the social influence occurs between people with different roles and with different levels of closeness to one another. We also note that although taking social influence from both peers and experts into consideration, our experiment still assumes a simplified version of the spread of misinformation on social media. In reality, users in social networks are organized in a more complex topology, making it possible for people to receive the same information multiple times from different paths, and receive several consistent or contradictory information simultaneously. It would be interesting and challenging to explore in the future how AI-based credibility indicator impacts people's perceptions of and engagement with the news under a nearly-realistic information spread scenario.

8 Conclusion

In this paper, we investigate into the effects of AI-based credibility indicators on people's perceptions of and engagement with online information, when these people are subject to social influence from both their laypeople peers and the experts. Via three randomized experiments, we show that despite of the social influence, the presence of accurate AI-based credibility indicators can help people determine the veracity of online information more accurately and reduce their propensity to engage with false information, and this is true regardless of whether the experts are self-claimed or verified. We also find that the effects of AI-based credibility indicators are particularly salient when AI predictions disagree with experts' opinions, and these effects are partially mediated through changing the peers' perceptions of news veracity. However, we also reveal that such effectiveness highly relies on the performance of the AI predictions, as people lack the capability to differentiate the correctness of AI predictions on news veracity, thus they can be easily misled by incorrect AI predictions. We hope this work could open more discussions on designing and evaluating interventions for mitigating misinformation in a world where people are subject to social influences from a crowd of mixed expertise.

Acknowledgments

We are grateful to the anonymous reviewers who provided many helpful comments. We thank the support of the National Science Foundation under grant IIS-2229876 and IIS-2340209 on this work. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors alone.

References

[1] Zhila Aghajari, Eric PS Baumer, and Dominic DiFranzo. 2023. What's the Norm Around Here? Individuals' Responses Can Mitigate the Effects of Misinformation Prevalence in Shaping Perceptions of a Community. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–27.

[2] Zhila Aghajari, Eric PS Baumer, Allison Lazard, Nabarun Dasgupta, and Dominic DiFranzo. 2024. Investigating the Mechanisms by which Prevalent Online Community Behaviors Influence Responses to Misinformation: Do Perceived Norms Really Act as a Mediator?. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–14.

[3] Meta AI. 2020. Here's how we're using AI to help detect misinformation. Retrieved October 10, 2022 from https://ai.facebook.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/

[4] Luca Maria Aiello, Martina Deplano, Rossano Schifanella, and Giancarlo Ruffo. 2012. People are strange when you're a stranger: Impact and influence of bots on social networks. In *Proceedings of the international AAAI conference on web and social media*, Vol. 6. 10–17.

[5] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 211–36.

[6] Hunt Allcott, Matthew Gentzkow, and Chuan Yu. 2019. Trends in the diffusion of misinformation on social media. *Research & Politics* 6, 2 (2019), 2053168019848554.

[7] Jack Andersen and Sille Obelitz Søe. 2020. Communicative actions we live by: The problem with fact-checking, tagging or flagging fake news–the case of Facebook. *European Journal of Communication* 35, 2 (2020), 126–139.

[8] Antonio A Arechar, Jennifer Allen, Adam J Berinsky, Rocky Cole, Ziv Epstein, Kiran Garimella, Andrew Gully, Jackson G Lu, Robert M Ross, Michael N Stagnaro, et al. 2023. Understanding and combatting misinformation across 16 countries on six continents. *Nature Human Behaviour* 7, 9 (2023), 1502–1513.

[9] John A Banas, Nicholas A Palomares, Adam S Richards, David M Keating, Nick Joyce, and Stephen A Rains. 2022. When Machine and Bandwagon Heuristics Compete: Understanding Users' Response to Conflicting AI and Crowdsourced Fact-Checking. *Human Communication Research* (2022).

[10] Melisa Basol, Jon Roozenbeek, and Sander Van der Linden. 2020. Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of cognition* 3, 1 (2020).

[11] Alessandro Bessi, Fabio Petroni, Michela Del Vicario, Fabiana Zollo, Aris Anagnostopoulos, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2015. Viral misinformation: The role of homophily and polarization. In *Proceedings of the 24th international conference on World Wide Web*. 355–356.

[12] Arnout B Boot, Katinka Dijkstra, and Rolf A Zwaan. 2021. The processing and evaluation of news content on social media is influenced by peer-user commentary. *Humanities and Social Sciences Communications* 8, 1 (2021), 1–11.

[13] Lia Bozarth, Jane Im, Christopher Quarles, and Ceren Budak. 2023. Wisdom of Two Crowds: Misinformation Moderation on Reddit and How to Improve this Process—A Case Study of COVID-19. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–33.

[14] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction* 5, CSCW1 (2021), 1–21.

[15] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2023. Are two heads better than one in ai-assisted decision making? comparing the behavior and performance of groups and individuals in human-ai collaborative recidivism risk assessment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[16] Alton YK Chua, Sin-Mei Cheah, Dion Hoe-Lian Goh, and Ee-Peng Lim. 2016. Collective rumor correction on the death hoax of a political figure in social media. AIS.

[17] Alicia Chung, Dorice Vieira, Tiffany Donley, Nicholas Tan, Girardin Jean-Louis, Kathleen Kiely Gouley, Azizi Seixas, et al. 2021. Adolescent peer influence on eating behaviors via social media: scoping review. *Journal of medical Internet research* 23, 6 (2021), e19697.

[18] Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, 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 (2020), 1073–1095.

[19] Jonas Colliander. 2019. "This is fake news": Investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media. *Computers in Human Behavior* 97 (2019), 202–215.

[20] John Cook, Stephan Lewandowsky, and Ullrich KH Ecker. 2017. Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PloS one* 12, 5 (2017), e0175799.

[21] Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. 2020. Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD*

international conference on knowledge discovery & data mining. 492–502.
[22] Shih-Chieh Dai, Yi-Li Hsu, Aiping Xiong, and Lun-Wei Ku. 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.
[23] Valdemar Danry, Pat Pataranutaporn, Matthew Groh, Ziv Epstein, and Pattie Maes. 2024. Deceptive AI systems that give explanations are more convincing than honest AI systems and can amplify belief in misinformation. *arXiv preprint arXiv:2408.00024* (2024).
[24] Ian Dennis. 2007. Halo effects in grading student projects. *Journal of Applied Psychology* 92, 4 (2007), 1169.
[25] MR DeVerna, HY Yan, KC Yang, and F Menczer. 2024. Fact-checking information from large language models can decrease headline discernment. *arXiv preprint arXiv:2308.10800* (2024).
[26] Chiara Patricia Drolsbach and Nicolas Pröllochs. 2023. Diffusion of community fact-checked misinformation on twitter. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–22.
[27] Ziv Epstein, Nicolo Foppiani, Sophie Hilgard, Sanjana Sharma, Elena Glassman, and David Rand. 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*, Vol. 16. 183–193.
[28] Ziv Epstein, Hause Lin, Gordon Pennycook, and David Rand. 2022. How many others have shared this? Experimentally investigating the effects of social cues on engagement, misinformation, and unpredictability on social media. *arXiv preprint arXiv:2207.07562* (2022).
[29] Ziv Epstein, Nathaniel Sirlin, Antonio Alonso Arechar, Gordon Pennycook, and David Rand. 2021. Social media sharing reduces truth discernment. (2021).
[30] Marina Ernst. 2024. Identifying textual disinformation using Large Language Models. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*. 453–456.
[31] Isabelle Freiling, Nicole M Krause, Dietram A Scheufele, and Dominique Brossard. 2021. Believing and sharing misinformation, fact-checks, and accurate information on social media: The role of anxiety during COVID-19. *New Media & Society* (2021), 14614448211011451.
[32] Noah E Friedkin and Eugene C Johnsen. 1990. Social influence and opinions. *Journal of Mathematical Sociology* 15, 3-4 (1990), 193–206.
[33] Werner Geyser. 2022. What is an influencer?–Social media influencers defined. *Influencer Marketing Hub* (2022).
[34] William Godel, Zeve Sanderson, Kevin Aslett, Jonathan Nagler, Richard Bonneau, Nathaniel Persily, and Joshua A Tucker. 2021. Moderating with the mob: Evaluating the efficacy of real-time crowdsourced fact-checking. *Journal of Online Trust and Safety* 1, 1 (2021).
[35] Lucas Graves and Federica Cherubini. 2016. The rise of fact-checking sites in Europe. *Digital News Project Report* (2016).
[36] Andrew Guess, Brendan Nyhan, and Jason Reifler. 2018. Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign. *European Research Council* 9, 3 (2018), 4.
[37] Douglas Guilbeault, Samuel Woolley, and Joshua Becker. 2021. Probabilistic social learning improves the public's judgments of news veracity. *Plos one* 16, 3 (2021), e0247487.
[38] Chen Guo, Nan Zheng, and Chengqi Guo. 2023. Seeing is not believing: a nuanced view of misinformation warning efficacy on video-sharing social media platforms. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–35.
[39] Hans WA Hanley, Deepak Kumar, and Zakir Durumeric. 2023. A Golden Age: Conspiracy Theories' Relationship with Misinformation Outlets, News Media, and the Wider Internet. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–33.
[40] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 1803–1812.
[41] Hyehyun Hong. 2013. Government websites and social media's influence on government-public relationships. *Public relations review* 39, 4 (2013), 346–356.
[42] Benjamin D Horne, Dorit Nevo, Sibel Adali, Lydia Manikonda, and Clare Arrington. 2020. Tailoring heuristics and timing AI interventions for supporting news veracity assessments. *Computers in Human Behavior Reports* 2 (2020), 100043.
[43] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* (2023).
[44] Emelia May Hughes, Renee Wang, Prerna Juneja, Tony W Li, Tanushree Mitra, and Amy X Zhang. 2024. Viblio: Introducing Credibility Signals and Citations to Video-Sharing Platforms. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.
[45] Farnaz Jahanbakhsh and David R Karger. 2024. A Browser Extension for in-place Signaling and Assessment of Misinformation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.
[46] Farnaz Jahanbakhsh, Yannis Katsis, Dakuo Wang, Lucian Popa, and Michael Muller. 2023. Exploring the use of personalized AI for identifying misinformation on social media. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–27.
[47] Farnaz Jahanbakhsh, Amy X Zhang, Adam J Berinsky, Gordon Pennycook, David G Rand, and David R Karger. 2021. Exploring lightweight interventions at posting time to reduce the sharing of misinformation on social media. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–42.
[48] Jennifer Jerit and Yangzi Zhao. 2020. Political misinformation. *Annual Review of Political Science* 23, 1 (2020), 77–94.
[49] Chenyan Jia, Alexander Boltz, Angie Zhang, Anqing Chen, and Min Kyung Lee. 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 (2022), 1–27.
[50] Prerna Juneja, Md Momen Bhuiyan, and Tanushree Mitra. 2023. Assessing enactment of content regulation policies: A post hoc crowd-sourced audit of election misinformation on YouTube. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–22.
[51] Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.
[52] Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang, and Soumendu Sinha. 2020. FNDNet–a deep convolutional neural network for fake news detection. *Cognitive Systems Research* 61 (2020), 32–44.
[53] Robert A Kaufman, Michael Robert Haupt, and Steven P Dow. 2022. Who's in the Crowd Matters: Cognitive Factors and Beliefs Predict Misinformation Assessment Accuracy. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–18.
[54] Nicole M Krause, Isabelle Freiling, Becca Beets, and Dominique Brossard. 2020. Fact-checking as risk communication: the multi-layered risk of misinformation in times of COVID-19. *Journal of Risk Research* 23, 7-8 (2020), 1052–1059.
[55] David La Barbera, Eddy Maddalena, Michael Soprano, Kevin Roitero, Gianluca Demartini, Davide Ceolin, Damiano Spina, and Stefano Mizzaro. 2024. Crowdsourced Fact-checking: Does It Actually Work? *Information Processing & Management* 61, 5 (2024), 103792.
[56] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q Vera Liao, and Chenhao Tan. 2023. Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1369–1385.
[57] Jiyoung Lee and Keeheon Lee. 2024. Measuring Falseness in News Articles based on Concealment and Overstatement. *arXiv preprint arXiv:2408.00156* (2024).
[58] Huaye Li and Yasuaki Sakamoto. 2013. The Influence of Collective Opinion on True-False Judgment and Information-Sharing Decision. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 35.
[59] Zhuoyan Li, Zhuoran Lu, and Ming Yin. 2023. Modeling human trust and reliance in ai-assisted decision making: A markovian approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 6056–6064.
[60] Gionnieve Lim and Simon T Perrault. 2024. Evaluation of an llm in identifying logical fallacies: A call for rigor when adopting llms in hci research. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*. 303–308.
[61] Houjiang Liu, Anubrata Das, Alexander Boltz, Didi Zhou, Daisy Pinaroc, Matthew Lease, and Min Kyung Lee. 2024. Human-centered NLP Fact-checking: Co-Designing with Fact-checkers using Matchmaking for AI. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–44.
[62] Yang Liu and Yi-Fang Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.
[63] Yang Liu and Yi-Fang Brook Wu. 2020. Fned: a deep network for fake news early detection on social media. *ACM Transactions on Information Systems (TOIS)* 38, 3 (2020), 1–33.
[64] Zhuoran Lu, Patrick Li, Weilong Wang, and Ming Yin. 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 (2022), 1–27.
[65] Zhuoran Lu, Dakuo Wang, and Ming Yin. 2024. Does more advice help? the effects of second opinions in AI-assisted decision making. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–31.
[66] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.
[67] Xiaoxiao Ma, Yuchen Zhang, Kaize Ding, Jian Yang, Jia Wu, and Hao Fan. 2024. On Fake News Detection with LLM Enhanced Semantics Mining. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 508–521.
[68] Jasmine E McNealy. 2023. All the Rumors are True: Verification, Actual Malice, and Celebrity Gossip. *Mo. L. Rev.* 88 (2023), 751.
[69] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27, 1 (2001),

415–444.
[70] Paul Mena. 2019. Principles and boundaries of fact-checking: Journalists' perceptions. *Journalism practice* 13, 6 (2019), 657–672.
[71] Paul Mena. 2020. Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. *Policy & internet* 12, 2 (2020), 165–183.
[72] Miriam J Metzger, Andrew J Flanagin, and Ryan B Medders. 2010. Social and heuristic approaches to credibility evaluation online. *Journal of communication* 60, 3 (2010), 413–439.
[73] Nicholas Micallef, Vivienne Armacost, Nasir Memon, and Sameer Patil. 2022. True or false: Studying the work practices of professional fact-checkers. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–44.
[74] Tanushree Mitra, Graham Wright, and Eric Gilbert. 2017. Credibility and the dynamics of collective attention. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–17.
[75] Ahmadreza Mosallanezhad, Mansooreh Karami, Kai Shu, Michelle V Mancenido, and Huan Liu. 2022. Domain Adaptive Fake News Detection via Reinforcement Learning. In *Proceedings of the ACM Web Conference 2022*. 3632–3640.
[76] Mehdi Moussaïd, Juliane E Kämmer, Pantelis P Analytis, and Hansjörg Neth. 2013. Social influence and the collective dynamics of opinion formation. *PloS one* 8, 11 (2013), e78433.
[77] Lev Muchnik, Sinan Aral, and Sean J Taylor. 2013. Social influence bias: A randomized experiment. *Science* 341, 6146 (2013), 647–651.
[78] Richard Nadeau, Edouard Cloutier, and J-H Guay. 1993. New evidence about the existence of a bandwagon effect in the opinion formation process. *International Political Science Review* 14, 2 (1993), 203–213.
[79] Stephen R Neely, Christina Eldredge, Robin Ersing, and Christa Remington. 2022. Vaccine hesitancy and exposure to misinformation: a survey analysis. *Journal of general internal medicine* (2022), 1–9.
[80] An T Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C Wallace, and Matthew Lease. 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.
[81] Melody Nouri. 2018. The power of influence: Traditional celebrity vs social media influencer. (2018).
[82] Folco Panizza, Piero Ronzani, Carlo Martini, Simone Mattavelli, Tiffany Morisseau, and Matteo Motterlini. 2022. Lateral reading and monetary incentives to spot disinformation about science. *Scientific reports* 12, 1 (2022), 1–15.
[83] Myrto Pantazi, Scott Hale, and Olivier Klein. 2021. Social and cognitive aspects of the vulnerability to political misinformation. *Political Psychology* 42 (2021), 267–304.
[84] Kunwoo Park, Haewoon Kwak, Hyunho Song, and Meeyoung Cha. 2020. "Trust Me, I Have a Ph. D.": A Propensity Score Analysis on the Halo Effect of Disclosing One's Offline Social Status in Online Communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 534–544.
[85] Gordon Pennycook, Adam Bear, Evan T Collins, and David G Rand. 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 (2020), 4944–4957.
[86] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 592, 7855 (2021), 590–595.
[87] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. 2020. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science* 31, 7 (2020), 770–780.
[88] Gordon Pennycook and David G Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116, 7 (2019), 2521–2526.
[89] Gordon Pennycook and David G Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188 (2019), 39–50.
[90] Gordon Pennycook and David G Rand. 2021. The psychology of fake news. *Trends in cognitive sciences* 25, 5 (2021), 388–402.
[91] David N Rapp and Nikita A Salovich. 2018. Can't we just disregard fake news? The consequences of exposure to inaccurate information. *Policy Insights from the Behavioral and Brain Sciences* 5, 2 (2018), 232–239.
[92] Julio CS Reis, André Correia, Fabricio Murai, Adriano Veloso, and Fabrício Benevenuto. 2019. Explainable machine learning for fake news detection. In *Proceedings of the 10th ACM conference on web science*. 17–26.
[93] Samuel C Rhodes. 2022. Filter bubbles, echo chambers, and fake news: How social media conditions individuals to be less critical of political misinformation. *Political Communication* 39, 1 (2022), 1–22.
[94] Jon Roozenbeek, Claudia R Schneider, Sarah Dryhurst, John Kerr, Alexandra LJ Freeman, Gabriel Recchia, Anne Marthe Van Der Bles, and Sander Van Der Linden. 2020. Susceptibility to misinformation about COVID-19 around the world. *Royal Society open science* 7, 10 (2020), 201199.
[95] Mohammed Saeed, Nicolas Traub, Maelle Nicolas, Gianluca Demartini, and Paolo Papotti. 2022. Crowdsourced fact-checking at Twitter: How does the crowd compare with experts?. In *Proceedings of the 31st ACM international conference on information & knowledge management*. 1736–1746.
[96] Vera Schmitt, Luis-Felipe Villa-Arenas, NIls Feldhus, Joachim Meyer, Robert P Spang, and Sebastian Möller. 2024. The Role of Explainability in Collaborative Human-AI Disinformation Detection. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2157–2174.
[97] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. 2024. Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.
[98] Connie Moon Sehat, Ryan Li, Peipei Nie, Tarunima Prabhakar, and Amy X Zhang. 2024. Misinformation as a harm: structured approaches for fact-checking prioritization. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–36.
[99] DongBack Seo and Jung Lee. 2014. Experts versus friends: To whom do i listen more? the factors that affect credibility of online information. In *International Conference on HCI in Business*. Springer, 245–256.
[100] Haeseung Seo, Sian Lee, Dongwon Lee, and Aiping Xiong. 2024. Reliability Matters: Exploring the Effect of AI Explanations on Misinformation Detection with a Warning. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 1395–1407.
[101] Haeseung Seo, Aiping Xiong, and Dongwon Lee. 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.
[102] Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. 2021. An exploratory study of COVID-19 misinformation on Twitter. *Online social networks and media* 22 (2021), 100104.
[103] Farhana Shahid, Shrirang Mare, and Aditya Vashistha. 2022. Examining source effects on perceptions of fake news in rural India. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–29.
[104] Lauren E Sherman, Ashley A Payton, Leanna M Hernandez, Patricia M Greenfield, and Mirella Dapretto. 2016. The power of the like in adolescence: Effects of peer influence on neural and behavioral responses to social media. *Psychological science* 27, 7 (2016), 1027–1035.
[105] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter* 19, 1 (2017), 22–36.
[106] Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining*. 312–320.
[107] Chenglei Si, Navita Goyal, Sherry Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé III, and Jordan Boyd-Graber. 2023. Large Language Models Help Humans Verify Truthfulness–Except When They Are Convincingly Wrong. *arXiv preprint arXiv:2310.12558* (2023).
[108] Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2021. The many dimensions of truthfulness: Crowdsourcing misinformation assessments on a multidimensional scale. *Information Processing & Management* 58, 6 (2021), 102710.
[109] Brian G Southwell and Emily A Thorson. 2015. The prevalence, consequence, and remedy of misinformation in mass media systems. 589–595 pages.
[110] Dominic Spohr. 2017. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business information review* 34, 3 (2017), 150–160.
[111] Marlis Stubenvoll, Raffael Heiss, and Jörg Matthes. 2021. Media trust under threat: Antecedents and consequences of misinformation perceptions on social media. *International Journal of Communication* 15 (2021), 22.
[112] Victor Suarez-Lledo and Javier Alvarez-Galvez. 2021. Prevalence of health misinformation on social media: systematic review. *Journal of medical Internet research* 23, 1 (2021), e17187.
[113] S Shyam Sundar. 2008. *The MAIN model: A heuristic approach to understanding technology effects on credibility*. MacArthur Foundation Digital Media and Learning Initiative Cambridge, MA.
[114] Briony Swire-Thompson, David Lazer, et al. 2020. Public health and online misinformation: challenges and recommendations. *Annu Rev Public Health* 41, 1 (2020), 433–451.
[115] Huiyun Tang, Gabriele Lenzini, Samuel Greiff, Björn Rohles, and Anastasia Sergeeva. 2024. "Who Knows? Maybe it Really Works": Analysing Users' Perceptions of Health Misinformation on Social Media. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 1499–1517.
[116] Haoheng Tang and Mrinalini Singha. 2024. A Mystery for You: A fact-checking game enhanced by large language models (LLMs) and a tangible interface. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–5.
[117] Petter Törnberg. 2018. Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS one* 13, 9 (2018), e0203958.
[118] Adaku Uchendu, Jooyoung Lee, Hua Shen, Thai Le, Dongwon Lee, et al. 2023. Does human collaboration enhance the accuracy of identifying llm-generated

deepfake texts?. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 11. 163–174.

[119] Sander Van Der Linden. 2022. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature medicine* 28, 3 (2022), 460–467.

[120] Sander Van der Linden, Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach. 2017. Inoculating the public against misinformation about climate change. *Global Challenges* 1, 2 (2017), 1600008.

[121] Federico Vegetti and Moreno Mancosu. 2020. The impact of political sophistication and motivated reasoning on misinformation. *Political Communication* 37, 5 (2020), 678–695.

[122] Rutvik Vijjali, Prathyush Potluri, Siddharth Kumar, and Sundeep Teki. 2020. Two stage transformer model for COVID-19 fake news detection and fact checking. *arXiv preprint arXiv:2011.13253* (2020).

[123] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[124] Christopher N Wahlheim, Timothy R Alexander, and Carson D Peske. 2020. Reminders of everyday misinformation statements can enhance memory for and beliefs in corrections of those statements in the short term. *Psychological Science* 31, 10 (2020), 1325–1339.

[125] Jinping Wang, Maria D Molina, and S Shyam Sundar. 2020. When expert recommendation contradicts peer opinion: Relative social influence of valence, group identity and artificial intelligence. *Computers in Human Behavior* 107 (2020), 106278.

[126] Yuping Wang, Chen Ling, and Gianluca Stringhini. 2023. Understanding the use of images to spread COVID-19 misinformation on Twitter. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–32.

[127] Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. 2019. Systematic literature review on the spread of health-related misinformation on social media. *Social science & medicine* 240 (2019), 112552.

[128] Ben Wasike. 2022. When the influencer says jump! How influencer signaling affects engagement with COVID-19 misinformation. *Social Science & Medicine* 315 (2022), 115497.

[129] Jevin D West and Carl T Bergstrom. 2021. Misinformation in and about science. *Proceedings of the National Academy of Sciences* 118, 15 (2021), e1912444117.

[130] Mika Westerlund. 2019. The emergence of deepfake technology: A review. *Technology innovation management review* 9, 11 (2019).

[131] Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter* 21, 2 (2019), 80–90.

[132] Jiechen Xu, Lei Han, Shazia Sadiq, and Gianluca Demartini. 2024. On the Role of Large Language Models in Crowdsourcing Misinformation Assessment. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 1674–1686.

[133] Ruichao Yang, Xiting Wang, Yiqiao Jin, Chaozhuo Li, Jianxun Lian, and Xing Xie. 2022. Reinforcement Subgraph Reasoning for Fake News Detection. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2253–2262.

[134] Waheeb Yaqub, Otari Kakhidze, Morgan L Brockman, Nasir Memon, and Sameer Patil. 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.

[135] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[136] Jingwen Zhang, Jieyu Ding Featherstone, Christopher Calabrese, and Magdalena Wojcieszak. 2021. Effects of fact-checking social media vaccine misinformation on attitudes toward vaccines. *Preventive Medicine* 145 (2021), 106408.

[137] Yang Zhang, Ruohan Zong, Lanyu Shang, Zhenrui Yue, Huimin Zeng, Yifan Liu, and Dong Wang. 2024. Tripartite Intelligence: Synergizing Deep Neural Network, Large Language Model, and Human Intelligence for Public Health Misinformation Detection (Archival Full Paper). In *Proceedings of the ACM Collective Intelligence Conference*. 63–75.

[138] Wayne Xin Zhao, Jing Liu, Yulan He, Chin-Yew Lin, and Ji-Rong Wen. 2016. A computational approach to measuring the correlation between expertise and social media influence for celebrities on microblogs. *World Wide Web* 19 (2016), 865–886.

[139] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.