The Effects of AI-based Credibility Indicators on the Detection and Spread of Misinformation under Social Influence

ZHUORAN LU, Department of Computer Science, Purdue University, USA
PATRICK LI, Department of Computer Science, Purdue University, USA
WEILONG WANG, Krannert School of Management, Purdue University, USA
MING YIN, Department of Computer Science, Purdue University, USA

Misinformation on social media has become a serious concern. Marking news stories with credibility indicators, possibly generated by an AI model, is one way to help people combat misinformation. In this paper, we report the results of two randomized experiments that aim to understand the effects of AI-based credibility indicators on people's perceptions of and engagement with the news, when people are under *social influence* such that their judgement of the news is influenced by other people. We find that the presence of AI-based credibility indicators nudges people into aligning their belief in the veracity of news with the AI model's prediction regardless of its correctness, thereby changing people's accuracy in detecting misinformation. However, AI-based credibility indicators show limited impacts on influencing people's engagement with either real news or fake news when social influence exists. Finally, it is shown that when social influence is present, the effects of AI-based credibility indicators on the detection and spread of misinformation are larger as compared to when social influence is absent, when these indicators are provided to people before they form their own judgements about the news. We conclude by providing implications for better utilizing AI to fight misinformation.

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

Additional Key Words and Phrases: misinformation, fake news, artificial intelligence, social influence, human-AI interaction

ACM Reference Format:
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. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 461 (November 2022), 27 pages. https://doi.org/10.1145/3555562

1 INTRODUCTION

In recent years, the growing problem of online misinformation (e.g., "fake news") has drawn much attention from society. Indeed, the wide and rapid spread of misinformation can cause confusion and panic among people and even misguide people's decisions in the real world, for example, regarding vaccination and voting [12, 13, 50]. To fight against online misinformation, one prevalent approach taken by social media platforms is to have third-party fact-checkers review online information and put warning labels on those content rated as false [24, 43]. However, the limited scalability

Authors' addresses: Zhuoran Lu, Department of Computer Science, Purdue University, West Lafayette, USA, lu800@purdue.edu; Patrick Li, Department of Computer Science, Purdue University, West Lafayette, USA, li3192@purdue.edu; Weilong Wang, Krannert School of Management, Purdue University, West Lafayette, USA, wang4167@purdue.edu; Ming Yin, Department of Computer Science, Purdue University, West Lafayette, USA, mingyin@purdue.edu.

2573-0142/2022/11-ART461
https://doi.org/10.1145/3555562

of manual fact-checking has prompted researchers and practitioners alike to explore alternative methods for signaling the credibility of online information, such as through automated AI-based technologies [28, 42]. To this end, a few recent studies show that AI-based credibility indicators, when presented in appropriate forms, can increase people's sensitivity in differentiating real and fake news [23, 44] and decrease people's propensity to share fake news [55].

Despite the promising empirical findings, most of the existing research examines the effectiveness of AI-based credibility indicators in a simplified setting in which people's judgements and actions on a piece of news are only decided by their own interpretation of the news and the influence brought up by the AI-based credibility indicators. On real-world social media platforms, however, people's perceptions of and engagement with the news are often shaped by *social influence* [49, 52], such as other people's opinions about the news as well. For example, on Twitter or Facebook, users can share a piece of news while adding their own comments, such as their explicit judgements on the veracity of the news, to it. As the news gets spread through the users' social connections, these comments can potentially impact how future users (who will see this news later) interpret the news. Therefore, a critical yet under-explored problem in deepening our understanding of AI-based credibility indicators is whether and how these indicators can assist people in identifying misinformation and stopping the spread of misinformation *when people are subject to social influence*.

There are reasons to conjecture the answer either way. On the one hand, as the news gets spread in the social network, if providing an AI-based credibility indicator nudges most individuals who see this news into aligning their belief in the veracity of the news with the AI model's predictions to some extent, the effects of AI-based credibility indicators on people's perceptions of and engagements with the news may not only still be present under the social influence, but even get amplified by social influence due to the possible occurrence of *information cascades* or *herding behavior* (i.e., people give up their own judgement and follow the crowd instead) [4, 47]. On the other hand, if some, or even the majority of people in the crowd disagree with the AI model in their evaluations of the credibility of the news after viewing the AI model's prediction, it may not be surprising to see the effects of AI-based credibility indicators disappear under the social influence, as people may start to question the trustworthiness of the AI model and consider the crowd's opinions as more informative. To further complicate matters, credibility indicators generated by AI models may not be perfectly accurate. Thus, it is unclear whether people, together with their peers, have the capability of effectively telling apart when the AI-based credibility indicators are reliable and when they are not, as well as how this capability may be affected by the ways that AI-based credibility indicators are presented such as *when* they are shown to people.

To thoroughly understand the effects of AI-based credibility indicators when people are consuming news under the social influence, in this paper, we conducted two pre-registered, randomized, human-subject experiments on Amazon Mechanical Turk (MTurk), recruiting subjects to review news stories. In our Experiment 1, we adapted the experimental setup from the classical information cascade experiment in economics [4] to simulate how people will be influenced by others' opinions in interpreting and reacting to a piece of news as the news gets diffused (i.e., subjects who received the news later could see the veracity judgements about the news that were made by all preceding subjects). Moreover, we created three treatments in Experiment 1 by varying *whether* AI-based credibility indicators were presented to subjects on the news and *when* they were presented (i.e., before/after subjects processed the news independently and got exposed to others' veracity judgements about the news). Thus, through Experiment 1, we aim to explore how the *presence* and the *timing* of an AI-based credibility indicator affect people's ability to detect misinformation as well as their willingness to share real and fake news when people are subject to social influence. Our Experiment 2 serves as a replication of Experiment 1, which allows us to examine the robustness of our experimental results under a more realistic setup (e.g., when the text of news is accompanied by an image), and it also enables us to compare the sizes of AI-based credibility indicators' effects on the detection and spread of misinformation, between the case when social influence is present and the case when social influence is absent.

The results of our two experiments consistently suggest that the presence of AI-based credibility indicators significantly increases people's tendency to align their veracity belief in a piece of news with the AI model's prediction, *regardless of* the correctness of the prediction. This means that when the credibility indicators provided by the AI model are correct (wrong), their presence will significantly improve (impair) both the ability of each individual and the collective ability of a group to accurately detect misinformation, even as people are subject to social influence when judging the veracity of the news. However, we find minimal evidence suggesting that these changes in people's ability to detect misinformation result in changes in people's sharing intention or the expected depth of spread for either real news or fake news—the only significant result we obtain is that providing correct AI-based credibility indicators increases an individual's willingness to share real news *relative to* fake news, but only when these indicators are presented to people *after* they have got the opportunity to process the news on their own. Finally, through comparisons of the effect sizes, we find that if the AI-based credibility indicators are presented to people *before* they process the news on their own, their impacts on the detection and spread of misinformation become larger when people are subject to social influence, compared to when people are not subject to social influence.

Together, our findings provide important implications on the possible benefits, risks, and limitations of utilizing AI technologies in combating online misinformation. Our findings also offer lessons on better leveraging AI-based credibility indicators to facilitate the detection of misinformation and reduce the spread of misinformation in a complex social environment. Lastly, our findings highlight the importance of studying misinformation in more realistic settings in controlled experiments. We conclude by discussing these implications.

2 RELATED WORK

2.1 Misinformation: Harms, Mechanisms, and Mitigation

In the past decade, misinformation has attracted research interests from a wide range of domains and angles. For instance, researchers have analyzed the harms brought by the spread of misinformation [2, 3, 25, 40], and found that misinformation can result in false perceptions and risky behavior, and it may even lead to a degree of distrust in authorized information. Unfortunately, researchers have found that false news cascades tend to diffuse to more people than the truth [50]. This observation raises the important question of understanding why people believe and are willing to share false news. Pennycook and Rand [39] synthesized the recent psychological literature that investigated into this question. Their review suggests that people's poor ability in discerning true and false information is associated with a lack of careful reasoning and relevant knowledge, as well as the use of various heuristics, while people's sharing of misinformation is more of an outcome of not paying attention to evaluate the accuracy of the information. Additional research has further shown that the spread of misinformation is exacerbated by people's selective exposure to information [10, 22, 46].

In light of the danger of misinformation, researchers and practitioners have explored ways to reduce the spread of misinformation [6, 26, 34, 37]. One of the most commonly adopted approaches is to have professional fact-checkers review online information and then display warning labels along with those inaccurate information spotted by the fact-checkers. It has been shown that the usage of such warning labels lowers people's perceived accuracy of misinformation [16, 51], reduces people's intention to share false news stories [30, 55], and improves people's ability in differentiating true and false information in the long term, especially when warning labels are provided as feedback after people have first processed the information independently [11]. The main drawback of such manual fact-checking approach is its limited scalability.

2.2 Utilize AI in Influencing Misinformation-Related Decision Making

In recent years, a rich set of empirical research has been conducted in understanding whether and how the presence of recommendations from an AI model can influence people's decision-making ability [9, 14, 27, 29, 41, 53, 56, 57]. In the context of misinformation research, to overcome the limited scalability of manual fact-checking, AI technologies have been developed to automate the detection of misinformation [19, 28, 31, 42, 54]. This opens up the possibility of supplying AI-based credibility indicators to people to assist them in evaluating news veracity. Most recently, researchers have started to conduct empirical research to examine the effectiveness of AI-based credibility indicators in influencing people's accuracy in evaluating news veracity and their news sharing intention. For example, Seo et al. [44] and Nguyen et al. [34] showed that displaying the AI model's recommendations on the veracity of news headlines/claims has the potential of increasing people's accuracy in detecting both real news and fake news, and they both emphasized the importance of the transparency of the model. Horne et al. [23] found that interventions from an AI model are more effective in increasing people's ability to detect fake news when the AI advice is tailored to confirmed heuristics used by news consumers. Moreover, through a comparative study, Yaqub et al. [55] found that AI-based credibility indicators can decrease people's propensity to share fake news, though their impacts are smaller than those of indicators provided by fact-checkers.

2.3 Social influence on Belief in Misinformation

The primary difference between our study and the earlier research that examines the effects of AI-based credibility indicators is that we consider a more realistic setting where people are influenced by others when evaluating the credibility of news, i.e., they are subject to *social influence*. Such social influence may come naturally from one's social networks [15, 48]—it can be as explicit as other's comments on whether a piece of news is fake or not, or as implicit as other's social engagement with the news (e.g., like or upvote a news post)—and it can potentially shape how people interpret and act upon the news. For example, it was found that seeing a comment from other people criticizing a news article as fake decreases one's likelihood of sharing it [17], while seeing a high level of social engagement statistics for a fake news story increases one's tendency to share it [5]. While these studies confirm that people's perceptions of news are affected by social influences, little is known on whether AI-based credibility indicators still have any impacts on people's detection and spread of misinformation, *when they are subject to social influence*. Our study, thus, fills this gap.

We note that in the real-world social media environment, as a piece of news diffuses to more people through a path in the social network, those ones who are exposed to the news later could be influenced by all the people who are exposed to the news earlier on the path. To simulate the *sequential* nature of how people get impacted by social influence, we get inspirations from the the information cascade literature in economics [1, 4, 21] when designing our experiment. Specifically, in the classical information cascade experiment in economics (e.g., [4]), there are two urns labeled A and B. For Urn A, the proportion of balls in it with a label of "a" is q ($q > 0.5$) while the proportion of balls with a label of "b" is $1 - q$. Conversely, for Urn B, the proportion of balls in it with a label of "b" is q while the proportion of balls with a label of "a" is $1 - q$. In the experiment, the experimenter will first randomly select a urn between Urn A and Urn B, with the probability of Urn A selected being p. Then, participants are asked to each draw a random ball from the selected urn in a sequential order and guess which urn (i.e., Urn A or Urn B) is selected by the experimenter, and participants are informed about the values of p and q. For the $(k+1)-$th participant in the sequence, they will first observe the label of the ball they draw, which are their "*private signal*s." Then, they will see the public decisions made by all k participants preceding them in terms of their guesses of which urn is selected. With all these information, they will finally make their *public decision*s on the selected urn. Empirically, it is observed that in such experiment, "*information cascade*" often occurs such that a participant will report a public decision that contradicts with the participant's private signal but aligns with public decisions made by previous participants (e.g., guess Urn A is selected despite the private signal is "b"), though such cascade does not necessarily lead to the correct decisions.

We adopted a similar setup in our experiment. However, our emphasis was not to understand whether information cascade occurs as people consume the news in a sequential order and be influenced by early receivers' opinions about the news when judging its veracity. Instead, our goal is to understand how providing AI-based credibility indicators impacts people's ability to detect misinformation and their willingness to share true and false information, in an environment where people are subject to social influence and a cascade of belief in news veracity *may* occur.

3 EXPERIMENT 1

In our Experiment 1, we aim to obtain an understanding of how providing an AI-based credibility indicator along with a piece of news will affect people's perception of the news and their intention to engage with the news, considering that these people are in a social environment such that their interpretation of the news is also affected by others' attitudes towards the news. Specifically, we are interested in examining how the *presence* and the *timing* of a credibility warning produced by an AI model will impact the detection and spread of misinformation. We ask:

- **RQ1**: How do the presence and the timing of an AI-based credibility indicator affect people's accuracy in detecting misinformation (i.e., fake news) when they are impacted by others' judgement on the veracity of the news?
- **RQ2**: How do the presence and the timing of an AI-based credibility indicator affect people's willingness to share real news and fake news when they are impacted by others' judgement on the veracity of the news?

To answer these research questions formally, we conducted a randomized, pre-registered experiment[1]. The design of our experiment was largely adapted from the classical information cascade experiments in economics [4], where human subjects were recruited from Amazon Mechanical Turk (MTurk) to review news while they could also observe the judgements on the news made by all preceding subjects who reviewed the news before themselves. In addition, subjects in different experimental treatments may or may not get access to the AI model's prediction of the credibility of the news. All of our experiments were approved by the IRB of the authors' institution.

3.1 Experimental Task

In this experiment, subjects were asked to complete a series of 10 tasks to review short news stories. News stories used in these tasks came from a dataset that we collected, which included a total of 40 news stories—20 true news stories (i.e., "real news") and 20 false news stories (i.e., "fake news")—related to COVID-19 (see the supplementary materials for the full list of news we used in this experiment). The veracity of the real news in our dataset was confirmed by cross-checking multiple reliable media outlets and peer-reviewed publications, while the fake news in our dataset was considered false because it was either disputed by authoritative sources (e.g., fact-checking sites) or conflicted with verified information. We decided to use COVID-19 related news in our experiment as COVID is a topic that has been intensively discussed by people today. Thus, having

[1] Our pre-registration document can be found here: https://aspredicted.org/u5fu5.pdf

subjects determine the veracity of COVID-19 related news and decide how much to engage with this news in our experiment has the potential to reflect people's real-world behavior, as they are likely interacting with massive amounts of COVID information in their real life.

In each task, the subject was presented with one piece of news that was randomly drawn from our news dataset and was asked to complete the following steps (see Figure 1 for an example of the task interface):

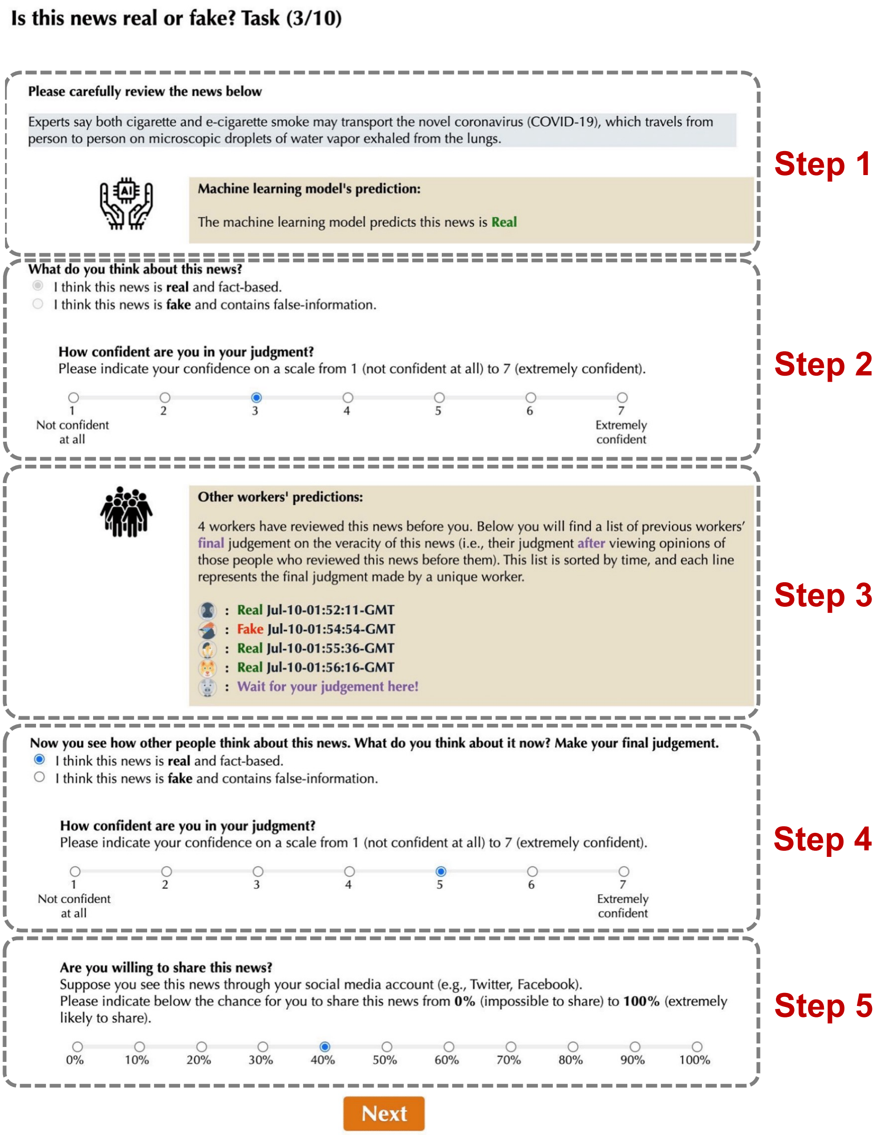
- **Step 1**: The subject was asked to carefully review the news.
- **Step 2**: The subject was asked to make an *initial* binary judgement on whether this news is real and fact-based or fake and contains false information. The subject also needed to report their confidence in this initial judgment on a 7-point Likert scale from 1 (not confident at all) to 7 (extremely confident).
- **Step 3**: If the subject was not the first one to review this news, they would be presented with a list of binary veracity judgments on this news that were made by other subjects who had reviewed this news before them, with the list sorted in chronological order.
- **Step 4**: After reviewing others' judgments, the subject needed to make their *final* binary judgment on the veracity of the news as well as reporting their confidence on the final judgement, again on a 7-point Likert scale. Note that for those subjects who would review this news after the current subject, in their Step 3, the veracity judgement they saw from the current subject would be this *final* judgement.
- **Step 5**: Finally, the subject was asked to indicate the likelihood for them to share the news on social media platforms as a percentage between 0% (impossible to share) to 100% (extremely likely to share).

In this 5-step procedure, we intended to use Step 3 to reflect real-world scenarios where people are exposed to "social influence" when interpreting a piece of news. Here in this experiment, we consider the social influence as other people's explicit judgement on the veracity of the news stories. Thus, given a particular piece of news, subjects who reviewed this news successively would produce a *sequence* of veracity judgements, and subjects who reviewed it at a later stage could see the veracity judgements made by all previous subjects. This design was adopted to simulate that as a piece of news gets spread through a path in the social network, late receivers of the news may be influenced by those early receivers in interpreting the news, thus a cascade of news veracity belief may occur. Note that one may map this 5-step procedure to the setup in the classical information cascade experiments [4] (see Section 2.3 for details)—An urn (i.e., the "real news" urn or the "fake news" urn) is selected in Step 1, and the subject observes their private signal (in Step 2) as well as the public decisions made by all preceding subjects (in Step 3), before they make their own public decision in Step 4.

3.2 Experimental Treatments

To reflect that in addition to being influenced by the opinions of "others," people may also get access to an AI model's prediction on news credibility and be influenced by it, we created three treatments in Experiment 1 by varying the presence and the timing of the AI-based credibility indicators:

- **Control**: Subjects in this treatment did *not* have access to the AI-based news credibility indicator in each task.
- **AI-before**: For subjects in this treatment, in each task, they saw the AI model's binary prediction on the credibility of the news in Step 1, which was *before* when they were asked to make their initial judgement on the veracity of the news and when they were exposed to other people's opinions about the news.



Is this news real or fake? Task (3/10)

Please carefully review the news below

Experts say both cigarette and e-cigarette smoke may transport the novel coronavirus (COVID-19), which travels from person to person on microscopic droplets of water vapor exhaled from the lungs.

Machine learning model's prediction:

The machine learning model predicts this news is **Real**

What do you think about this news?

● I think this news is **real** and fact-based.
○ I think this news is **fake** and contains false-information.

How confident are you in your judgment?
Please indicate your confidence on a scale from 1 (not confident at all) to 7 (extremely confident).

○ 1 Not confident at all ○ 2 ● 3 ○ 4 ○ 5 ○ 6 ○ 7 Extremely confident

Other workers' predictions:

4 workers have reviewed this news before you. Below you will find a list of previous workers' final judgement on the veracity of this news (i.e., their judgment after viewing opinions of those people who reviewed this news before them). This list is sorted by time, and each line represents the final judgment made by a unique worker.

: **Real Jul-10-01:52:11-GMT**
: **Fake Jul-10-01:54:54-GMT**
: **Real Jul-10-01:55:36-GMT**
: **Real Jul-10-01:56:16-GMT**
: **Wait for your judgement here!**

Now you see how other people think about this news. What do you think about it now? Make your final judgement.

● I think this news is **real** and fact-based.
○ I think this news is **fake** and contains false-information.

How confident are you in your judgment?
Please indicate your confidence on a scale from 1 (not confident at all) to 7 (extremely confident).

○ 1 Not confident at all ○ 2 ○ 3 ○ 4 ● 5 ○ 6 ○ 7 Extremely confident

Are you willing to share this news?
Suppose you see this news through your social media account (e.g., Twitter, Facebook).
Please indicate below the chance for you to share this news from **0%** (impossible to share) to **100%** (extremely likely to share).

○ 0% ○ 10% ○ 20% ○ 30% ● 40% ○ 50% ○ 60% ○ 70% ○ 80% ○ 90% ○ 100%

Fig. 1. An example of our task interface (for the AI-BEFORE treatment, so the model's prediction on news credibility is shown in Step 1). All steps are displayed on the same page, but are shown to subjects progressively.

- **AI-after**: For subjects in this treatment, in each task, they saw the AI model's binary prediction on the credibility of the news in Step 3 along with previous subjects' judgements on the veracity of the news, which was *after* when they were asked to make their initial veracity judgement, while they were exposed to other people's opinions about the news.

The AI-BEFORE treatment was designed to simulate the scenario that social media platforms directly display AI-based credibility warning labels along with each news item, so that these labels have the potential to shape people's belief about the news even before they form their own opinions independently. This is similar to how Twitter and Facebook have applied fact-checking labels to

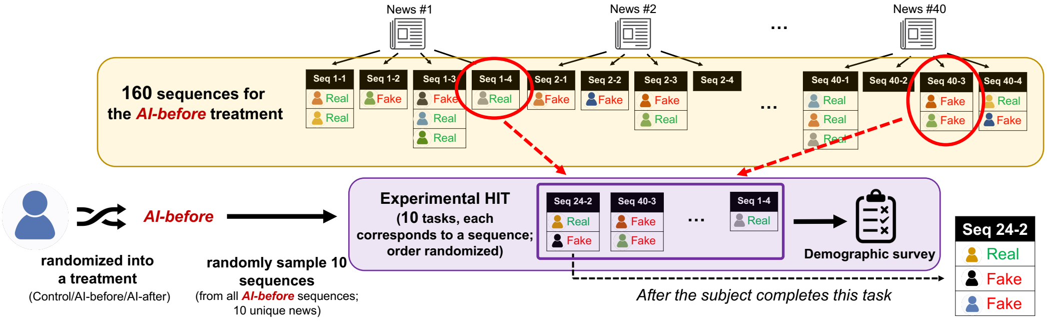


Fig. 2. The procedure of our Experiment 1 (the case that a subject gets randomly assigned to the AI-before treatment is shown).

content on their platforms today [24, 43]. Meanwhile, the AI-after treatment could reflect the scenario that social media platforms simply provide the AI model's prediction on the veracity of a news item to people as a reference (e.g., as a third-party plugin) along with others' opinions about the news, after people are actively encouraged to consider their own opinions about the news.

Importantly, in our experiment, the AI model we used in the AI-before and AI-after treatments was the *same*—using a subset of the COVID-19 healthcare misinformation dataset introduced in [18] as the training dataset, we developed a multinomial naive Bayes model to classify the veracity of COVID news. On the 40 news stories that we included in our experiment, the accuracy of our AI model was 75%, and the false positive rate and false negative rate of the model were both 25% (i.e., the AI model's accuracy was 75% for both the real news and the fake news in our dataset). Note that we tuned the AI model to achieve a 75% accuracy on our news dataset because this allows us to get sufficient data to examine the effects of AI-based credibility indicators both for the case that the model's prediction is *right* and for the case that the model's prediction is *wrong*. Whether the AI model's prediction on news credibility is correct or not is *not* communicated to subjects anytime during the experiment, though.

3.3 Experiment Procedure

We posted our experiment as a HIT on Amazon Mechanical Turk (MTurk) to U.S. workers only, and we allowed each worker to take our experiment HIT at most once. Figure 2 provides a schematic illustration of our experimental procedure.

Specifically, upon subjects' arrival, we randomly assigned each of them to one of the three experimental treatments. Subjects were told that in this HIT, they would complete 10 tasks to review 10 pieces of COVID-19 related news and determine the veracity of each news, together with other MTurk workers. Following the design of the information cascade experiments [4], we informed subjects that the 10 pieces of news that they would review in the HIT would be randomly drawn from a news dataset, with half of the news in this dataset being real (i.e., fact-based) and the other half being fake (i.e., contains false information), so that all subjects had a common prior belief of the veracity distribution of the news[2]. We then presented detailed instructions to subjects on what they would need to do in each task (e.g., the 5 steps as discussed in Section 3.1; the availability of the AI-based credibility indicators was also communicated to subjects if they were assigned to the AI-before or AI-after treatment). After that, subjects were asked to complete two more

[2] In a real-world social media environment, such information is unlikely available to people, and different individuals may also have different prior beliefs. We thus examined the robustness of our experimental results in Experiment 2 for the case when such information is not provided to subjects.

steps before proceeding to the actual experiment: (1) they needed to complete a consent form; (2) they needed to select an avatar to represent themselves in our experiment, so that their veracity judgement on a piece of news could be shown along with their chosen avatars to later subjects.

As the subject entered the actual experiment, we selected the 10 tasks for the subject to complete in the HIT. Each task was characterized by a piece of news, as well as a sequence of veracity judgements about this news sorted in the temporal order, which reflected whether those subjects who had reviewed this news subsequently thus far believed it as real or fake. Then, in each task, the subject followed the 5-step procedure as described before to review the news and others' judgements on it, evaluate its veracity, and determine their willingness to engage with it. Once they completed a task, their final veracity judgement on the news in that task (produced in Step 4) would be added to the end of the judgement sequence (see the bottom part of Figure 2).

Note that in practise, within each treatment, we formed 4 judgement sequences for each of the 40 news in our dataset to simulate that each news gets spread in the social network through 4 paths[3]. That is, given a specific experimental treatment X (e.g., AI-BEFORE), we had $40 \times 4 = 160$ judgement sequences—"Sequence $i - j$" ($1 \leq i \leq 40, 1 \leq j \leq 4$) represented the j-th sequence of veracity judgements on news i, which were made by a group of subjects who were all assigned to treatment X and reviewed this news following a sequential order (see the top part of Figure 2). Thus, for a subject of treatment X, the 10 tasks that they worked on were randomly selected based on the following procedure: (1) Firstly, we randomly selected a set of 10 unique news $\{n_1, n_2, \cdots, n_{10}\}$ from the 40 news in our dataset; (2) Secondly, for each selected news $n_t (1 \leq t \leq 10)$, we randomly selected a judgement sequence "Sequence $n_t - j_t$" ($1 \leq j_t \leq 4$) from the 4 sequences of news n_t in treatment X. In other words, in this subject's t-th task in the HIT, they would review news n_t, as well as the veracity judgements made by all previous subjects in "Sequence $n_t - j_t$" of treatment X, before they made their final judgement on the news veracity. This task selection design helps us ensure that all veracity judgements in the same sequence were made under the same condition, with respect to whether and when subjects got access to the AI-based credibility indicators.

To minimize the chance that subjects in our experiment were misled by the fake news that they reviewed in our experiment, we debriefed subjects about the ground-truth veracity label for each news that they reviewed after they completed all tasks in the HIT. To filter out potential spammers, we also included an attention check question[4] in our HIT, and workers could complete the HIT only if they passed the attention check. Finally, we asked subjects to fill out a brief demographic survey (e.g., age, gender) before they submitted the HIT.

The base payment of the HIT was \$1.2. To encourage subjects to carefully analyze the veracity of news in each task, we told subjects that they could earn a bonus of 5 cents for each correct final veracity judgment that they made, if the accuracies of their final veracity judgments in our HIT were over 60%. Thus, the subject could receive a bonus payment up to \$0.5 in this HIT.

3.4 Analysis Methods

3.4.1 Independent Variables. The main independent variable we used in our analysis is the experimental treatment that a subject was assigned to, i.e., the presence and the timing of the AI-based credibility indicators.

3.4.2 Dependent Variables. To understand how the presence and the timing of the AI-based credibility indicators change people's accuracy in detecting misinformation (**RQ1**), we pre-registered two dependent variables: (1) the accuracy of a subject's *final* judgement on the news veracity (i.e.,

[3] We did so to ensure an adequate sample size for our experiment.

[4] In the attention check question, subjects were asked to determine whether the statement "Washington DC is the capital city of the USA" is real or fake.

"individual-level accuracy"), and (2) the accuracy of the *majority* final veracity judgement made by subjects in a sequence on the news of that sequence (i.e., "sequence-level accuracy")[5]. The first dependent variable enables us to take a microscopic look at the effects of AI-based credibility indicators on the capability of each *individual* in detecting misinformation. However, the change in some individuals' capability in detecting misinformation does not necessarily imply a change in the crowd's capability in detecting misinformation as a *group*. Thus, we defined a group of people's veracity perception of a piece of news as the majority judgement made by all members of the group, so the second dependent variable allows us to take a macroscopic perspective and examine how the AI-based credibility indicators influence the collective capability of a group of people in identifying misinformation as they process the news content sequentially. Besides, by considering the majority final veracity judgement made by the first n subjects in each sequence, we can also obtain a more fine-grained understanding of how the crowd's capability in detecting misinformation changes as the size of the group n gets larger (i.e., the news spreads deeper).

Moreover, to understand how the presence and the timing of the AI-based credibility indicators change people's willingness to spread the news (**RQ2**), we also considered a few dependent variables. The first one was a subject's average level of self-reported willingness to share real (or fake) news. Earlier research has confirmed the validity of such self-reported sharing intention measures as they are found to be correlated with people's actual sharing behavior on social media platforms [33]. As the second dependent variable, we computed the expected depth of the spread for each real (or fake) news using the sharing willingness data we collected from our subjects. In particular, given a sequence formed for a piece of real (or fake) news, we simulated each subject's sharing decision based on the sharing likelihood that the subject reported (e.g., if a subject's willingness to share the news was 40%, then with a probability of 40%, this subject's simulated decision was "share"), and we defined the depth of the spread as the maximum number of consecutive subjects in the sequence whose simulated decision was "share" starting from the first subject in the sequence (i.e., the news stopped spreading once it encountered the first "not share" decision in the sequence). The *expected* depth of the spread for the news in this sequence was then computed as the average depth value across 1000 such simulations. Ideally, we hope people could minimize their sharing of fake news as much as possible (i.e., lower sharing intention and smaller expected depth of spread for fake news) so that misinformation would not be propagated to and affect a bigger crowd.

3.4.3 Statistical methods. For **RQ1**, we conducted the one-way analysis of variance (ANOVA) across the three treatments on the individual-level accuracy, and post-hoc Tukey's test was used to detect significant differences between pairs of treatments. We further visualized how the sequence-level accuracy for the three treatments changes as the length of the sequence gets longer. Since the AI model had an accuracy of 75% on the news that we included in our experiment, we further separated the news into two subsets based on whether the model made a correct prediction on its veracity or not and repeated the analyses above on these two subsets.

For **RQ2**, we conducted ANOVA on individual subjects' sharing intention for a piece of news and the expected depth of spread of the news for both the set of real news and the set of fake news. Similar to that for **RQ1**, Tukey's test was used for the post-hoc pairwise comparisons, and we again repeated all of these analyses for the two subsets of news where the AI model was correct or wrong separately.

[5] If there was no majority judgement among subjects' final judgements in a sequence, we randomly picked a label (i.e., real or fake) to break the tie.

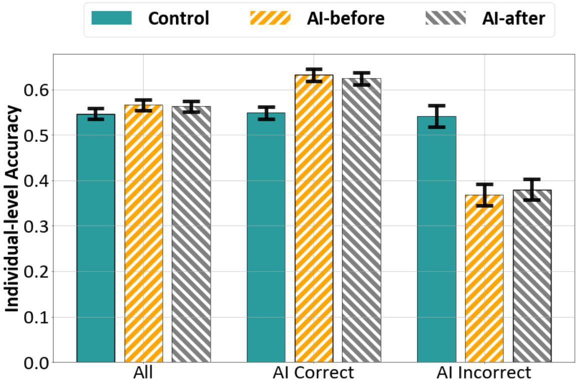


Fig. 3. The impacts of AI-based credibility indicators on individual subject's final accuracy in judging news veracity. Error bars represent the standard errors of the mean.

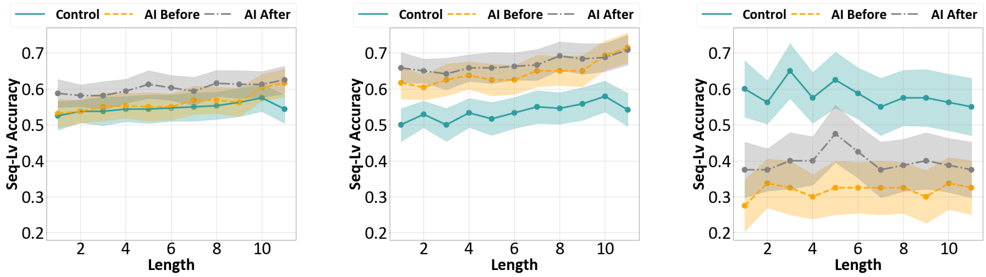
3.5 Results

In total, 538 subjects (34.3% female, with the majority aged between 25 and 44) took our experiment HIT and passed the attention check. As a result, for each of the 160 × 3 = 480 sequences we formed in the experiment, we had at least 9 subjects, and most of the sequences contained 11 subjects. In the following, we analyzed these experimental data to answer our research questions.

3.5.1 The Effects on Detection of Misinformation. We set out to analyze the effects of AI-based credibility indicators on people's ability in detecting misinformation when they are also influenced by others' opinions in judging the veracity of news (i.e., **RQ1**).

When people are under social influence, AI-based credibility indicators can still change an individual's accuracy in detecting misinformation, *regardless of their correctness.* As shown in Figure 3, when examining the individual subject's accuracy of *final* veracity judgements that were made on all news stories (i.e., the "All" group in Figure 3), the impact of the AI-based credibility indicators is not clear. However, when we focus on the cases where the AI model's prediction on the veracity of the news is *correct* (the "AI Correct" group), we find that a correct AI model prediction helps people increase their accuracy in judging the veracity of news, despite they are influenced by others' opinions when making such judgements. Indeed, a one-way ANOVA confirms that the differences across the three treatments on the individual-level accuracy are significant (F(2,3958) = 11.79, $p < 0.001$) when the AI model is correct. Post-hoc pairwise comparisons further show that these significant differences are mainly caused by the *presence* of the AI-based credibility indicators (Control vs. AI-before: $p = 0.001$, Cohen's $d = 0.17$; Control vs. AI-after: $p = 0.001$, Cohen's $d = 0.15$), while the timing of when the AI-based credibility indicators are displayed does not lead to substantially different impacts. On the contrary, when restricting our attention to the cases where the AI model's prediction on the veracity of the news is *incorrect* (the "AI Incorrect" group), we see exactly the opposite—providing a wrong AI model prediction results in a significant decrease in subjects' final veracity judgement accuracy (F(2,1317) = 17.16, $p < 0.001$). Again, only the presence, but not the timing, of the AI-based credibility indicators leads to the significant differences across the three treatments.

Together, these observations imply that providing AI-based credibility indicators to people has the effect of swaying their belief on the news veracity to be *in favor of the AI model's prediction*, thus people's veracity judgement accuracy changes accordingly with the AI model's correctness.



(a) Seq-level accuracy (All) (b) Seq-level accuracy (AI Correct) (c) Seq-level accuracy (AI Incorrect)

Fig. 4. The impacts of AI-based credibility indicators on the crowd's accuracy in judging news veracity when considering all news (4a), only the news on which the AI model's prediction is correct (4b), or only the news on which the AI model's prediction is wrong (4c). Error shades represent the standard errors of the mean.

AI-based credibility indicators affect the crowd's accuracy in detecting misinformation as the news spreads in the crowd, though the effect size does *not* change with the depth of spread of the news. Figure 4 compares the sequence-level accuracy across the three treatments when varying the "length" of the sequence (i.e., considering only the first n subjects in each sequence, $n = 1, 2, \cdots, 11$), which effectively reflects the crowd's ability in detecting misinformation as the news spreads deeper. Similar to the observations that we've made on the individual level, here, we again find that when the AI model provides a correct (incorrect) prediction on the veracity of a piece of news, the crowd's accuracy in judging the news veracity is increased (decreased)—for example, when considering all subjects in each sequence (i.e., $n = 11$), the sequence-level accuracy in the AI-before and AI-after treatments is significantly higher than that in the Control treatment when the credibility warning given by the AI model is correct ($F(2, 476) = 5.19, p = 0.006$). The specific timing of when the credibility warnings are displayed, again, does not seem to have a clear impact on the sequence-level accuracy. A closer look at Figure 4 further indicates that the effect of AI-based credibility indicators on the crowd's veracity judgement accuracy is neither amplified nor attenuated as the news gets spread to more people and the size of the crowd increases (i.e., n becomes larger).

3.5.2 The Effects on Spread of Misinformation. We now move on to examine the effects of AI-based credibility indicators on people's willingness to share real and fake news when they are influenced by others' opinions in interpreting these news (i.e., **RQ2**).

AI-based credibility indicators have *no* significant impacts on an individual's sharing intention on real news or fake news. First, we compare a subject's self-reported sharing intention on both real news and fake news across the three treatments. The results suggest that when people are under the social influence in interpreting the news, displaying the AI-based credibility warnings with an accuracy of 75% along with each news does *not* change people's intention to share either real news or fake news, and this is true even after we analyze the cases that the AI model makes correct/wrong veracity predictions separately. In particular, when the AI model provides correct prediction on the veracity of a piece of news, subjects slightly increase their willingness to share real news (AI-before − Control: $\Delta M = 1.42$; AI-after − Control: $\Delta M = 1.05$) and decrease their willingness to share fake news (AI-before − Control: $\Delta M = -1.00$; AI-after − Control: $\Delta M = -3.09$), although these differences are not significant ($p > 0.05$). On the other hand, when the AI model's prediction is wrong, the incorrect credibility indicators nudge people into sharing less real news (AI-before − Control: $\Delta M = -4.35$; AI-after − Control:

$\Delta M = -6.42$) and more fake news (AI-before − Control: $\Delta M = 2.37$; AI-after − Control: $\Delta M = 3.55$). Again, these differences are not statistically significant.

AI-based credibility indicators have *no* significant impacts on how deep real or fake news spreads. We obtain similar findings when examining the impacts of AI-based credibility indicators on the expected depth of spread for both real news and fake news. For example, when the credibility warnings provided by the AI model are correct, compared to that in the Control treatment, the expected depth of spread for real news becomes larger in both treatments where subjects had access to the AI model's veracity prediction (AI-before − Control: $\Delta M = 0.26$, AI-after − Control: $\Delta M = 0.28$), while the expected depth of spread for fake news only becomes slightly smaller in the AI-after treatment (AI-before − Control: $\Delta M = 0.20$, AI-after − Control: $\Delta M = -0.06$). Conversely, when the AI model makes wrong predictions on the veracity of news, real news tends to be propagated to fewer people (AI-before − Control: $\Delta M = -0.11$, AI-after − Control: $\Delta M = -0.04$) while fake news tends to be propagated to more people (AI-before − Control: $\Delta M = 0.12$, AI-after − Control: $\Delta M = 0.07$), compared to when the AI model's prediction is absent. Nevertheless, results of ANOVA tests show that *none* of these differences are significant at the level of $p = 0.05$.

3.5.3 Exploratory Analysis. So far, our findings suggest that even as people are influenced by others' opinions in interpreting the news, providing AI-based credibility indicators along with news items tends to encourage people to *align* their veracity belief about the news with the AI model's prediction. To obtain deeper insights into the mechanisms underlying the AI-based credibility indicators' impacts on people under the social influence, we conduct a set of exploratory analyses.

Providing AI-based credibility indicators *before* people process news independently significantly shapes people's first impression of the news. Using a one-way ANOVA test, we detect a significant difference across the three treatments on the likelihood of a subject's *initial* veracity judgement to be the *same* as the AI model's prediction ($F(2, 5275) = 36.78, p < 0.001$)[6]. In particular, compared to subjects in the Control or AI-after treatments, subjects in the AI-before treatment were significantly more likely to align their own initial veracity judgement of the news with the AI model's predictions (AI-before vs. Control: $p = 0.001$, Cohen's $d = 0.24$; AI-before vs. AI-after: $p = 0.001$, Cohen's $d = 0.26$). Furthermore, we compute subjects' initial confidence in the AI model's veracity prediction on a piece of news based on their self-reported confidence in their own initial veracity judgements—if the subjects' initial judgements were the same as the model's prediction, their initial confidence in the model's prediction would be the same as their confidence in their initial judgements; otherwise, we denote their initial confidence in the model's prediction as the opposite of their confidence in their initial judgement. Doing so, we find that subjects in the AI-before treatment were significantly more confident in the AI model's prediction than subjects in the other two treatments when making their initial veracity judgements ($p = 0.001$ for both comparisons). In other words, a key reason that explains the effectiveness of presenting AI-based credibility indicators *before* people process the news independently is that it frames people's first impression of the news by nudging them into confidently believing in the AI model's veracity prediction on the news.

People increase their alignment with the AI model more when the model's prediction is *different* from their initial veracity judgements, especially when the AI-based credibility indicators are displayed *after* people process the news independently. Next, we aim to obtain

[6] For subjects in the Control treatment, even though they never saw the AI model's prediction, we knew what our AI model would have predicted for each news story. We thus used this "hypothetical" AI model prediction to determine whether the subject's initial judgement was the same as the model.

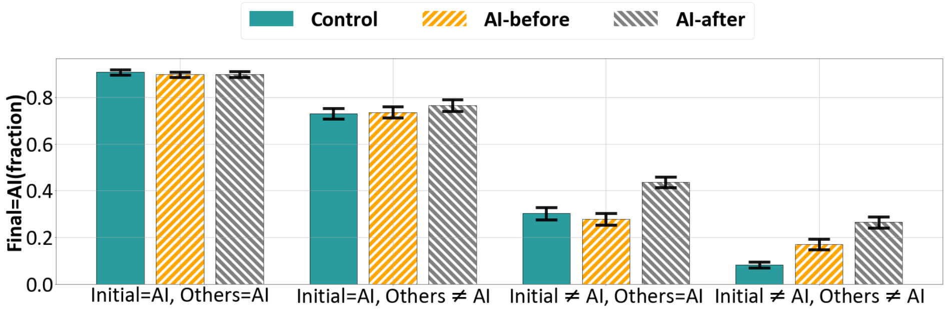


Fig. 5. The likelihood of a subject's final veracity judgment on a piece of news being the same as the AI model's prediction, when separating the data into four cases based on whether the subject's initial judgement was the same as the AI model, and whether the majority judgement of other people (i.e., subjects who reviewed this news before the current subject) was the same as the AI model. Error bars represent the standard errors of the mean.

a deeper understanding of whether and how providing AI-based credibility indicators along with a news item changes the ways that individuals updates their *final* veracity judgements on the news after they are exposed to social influence. In particular, consider a subject who needs to make the final veracity judgement on a news item. Depending on whether the subject's initial veracity judgement is the same as the AI model's prediction, and whether the majority judgement made by the set of subjects who have seen this news *so far* (i.e., the majority judgement of "others") is the same as the AI model's prediction, we have 4 possible scenarios in total. Figure 5 compares the likelihood of a subject's final veracity judgement on a news item being the same as the AI model's prediction across the three experimental treatments for each of these 4 scenarios.

We start by examining the two scenarios where the subject's initial veracity judgement is the *same* as the AI model's prediction (i.e., the "Initial=AI, Others=AI" and "Initial=AI, Others≠AI" group in Figure 5). Under both scenarios, we find that there are *no* significant differences across the three treatments on how likely the subject's final veracity judgement would be the same as the AI model. This means that seeing the AI model supporting their own judgement (i.e., "Initial=AI") has limited impacts on strengthening people's belief in their own judgement, regardless of whether others agree with them or not. For both scenarios, we also find that people's final confidence in the AI model's prediction[7] is statistically the same across the three treatments, implying that the agreement between the AI model and one's initial veracity judgement does not boost people's confidence in their own judgement when they are subject to social influence.

On the contrary, for the two scenarios where the subject's initial veracity judgement is *different* from the AI model's prediction (i.e., the "Initial≠AI, Others=AI" and "Initial≠AI, Others≠AI" group in Figure 5), we find that there are substantial differences across the three treatments on how likely the subject would eventually "switch" to align with the model's prediction. For example, when people disagree with others in their judgement on the veracity of a news item and the AI model supports the crowd (i.e., the "Initial≠AI, Others=AI" scenario), we find that subjects in the AI-AFTER treatment were significantly more likely to switch to the AI model's prediction than subjects in the other treatments (CONTROL vs. AI-AFTER: $p = 0.001$, AI-BEFORE vs. AI-AFTER: $p < 0.001$), and their final confidence in the AI model's prediction was also significantly higher (CONTROL vs. AI-AFTER: $p = 0.002$, AI-BEFORE vs. AI-AFTER: $p = 0.001$). Moreover, when people find the

[7] The subjects' final confidence in the AI model's prediction are computed based on their self-reported confidence in their own final veracity judgements, similar to how the subjects' initial confidence in the AI model's prediction are computed.

AI model's prediction to be different than both the crowd and themselves (i.e., the "Initial≠AI, Others≠AI" scenario), we find that those in the AI-after treatment were more likely to eventually switch to the AI model's prediction and have higher final confidence in the AI model's prediction than those in the AI-before treatment (p = 0.001 for final agreement with AI and p = 0.008 for final confidence in AI), who in turn were more likely to align their final judgement with the AI model than those in the Control treatment (p = 0.004 for final agreement with AI and p = 0.008 for final confidence in AI). We conjecture that a plausible explanation for these observations is that when people see the AI model makes a different veracity prediction than themselves, they treat it as a sign of the AI having access to additional information that they don't have, which leads to the increase in their likelihood of aligning with the AI model. Further, such an increase may be less salient for subjects in the AI-before treatments, possibly because they have a strong subjective belief in their own initial veracity judgement—after all, they were the ones who had actively chosen a different initial judgement than the AI even after seeing the AI model's prediction.

4 EXPERIMENT 2

In Experiment 1, we find that when people are subject to social influence, the presence of the AI-based credibility indicator still has a significant effect on people's capability to detect misinformation, though such impact does not seem to directly result in a significant change on people's willingness to engage with the news. We note that as the design of Experiment 1 was directly adapted from the information cascade experiments in economics, some aspects of the design might not perfectly reflect the realistic social media environment—for example, in the real world, people may have little or different knowledge about the distribution of the veracity of news, and the news items are also often displayed together with non-textual components (e.g., images, videos) which may influence people's perceptions of the news.

Thus, in our Experiment 2, we aim to answer **RQ1** and **RQ2** again under a more realistic experimental setting to examine the robustness of our Experiment 1 results. In addition, to put the effects of AI-based credibility indicators on the detection and spread of misinformation that we've observed in our experiment into context, an interesting question to ask is how do the size of these effects—which are obtained when social influence is *present*—compare with the effect size in those cases that social influence is *absent*. We thus ask an additional research question in Experiment 2:

- **RQ3**: Compared to when social influence doesn't exist, how do the magnitude of the effects of AI-based credibility indicators on people's accuracy in detecting misinformation and people's willingness to share real or fake news change when social influence is present?

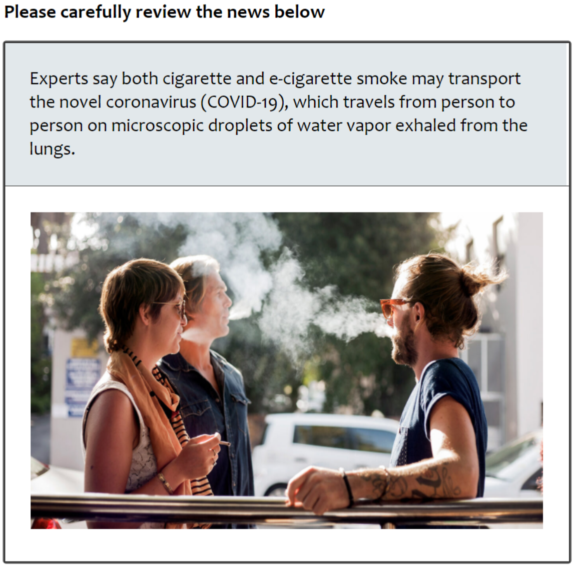
To answer these questions, we conducted a second randomized, pre-registered human-subject experiment on Amazon Mechanical Turk (MTurk)[8].

4.1 Experimental Design

4.1.1 Experiment Task. In Experiment 2, we again asked subjects to complete a series of 10 tasks to review COVID-19 related news, which were randomly sampled from the same dataset that we had used in Experiment 1. To reflect the real-world formats of news stories, in this experiment, the news we showed to subjects consisted of not only the text but also an image. Figure 6 shows an example of the news that we showed to subjects in Experiment 2. The steps that a subject needed to follow in each task will be detailed next in Section 4.1.2.

4.1.2 Experiment Treatments. To allow us to answer **RQ3**, in this experiment, we adopted a 2 × 3 factorial design. The first factor we varied was *the existence of social influence*:

[8] Our pre-registration document can be found here: https://aspredicted.org/rh32h.pdf.



Please carefully review the news below

Experts say both cigarette and e-cigarette smoke may transport the novel coronavirus (COVID-19), which travels from person to person on microscopic droplets of water vapor exhaled from the lungs.

Fig. 6. An example of the news we showed to subjects in Experiment 2.

- **No social influence (Independent)**: Subjects in this treatment reviewed the news *independently* without observing the veracity judgments made by other people who previously reviewed the same news.
- **With social influence (Non-Independent)**: Subjects in this treatment were *subject to social influence* when reviewing the news; that is, they could see the veracity judgments made by other people who previously reviewed the same news.

The second factor that we varied was *the existence and timing of the AI-based credibility indicators*. As that in Experiment 1, we included three levels: Control, AI-before and AI-after. Specifically, for subjects who were exposed to social influence (i.e., subjects in the Non-Independent treatments), if they were also assigned to the Control (alternatively, AI-before or AI-after) treatment in Experiment 2, in each task they would need to follow exactly the same 5-step procedure as those subjects in the Control (alternatively, AI-before or AI-after) treatment in Experiment 1. In contrast, for subjects who were not exposed to social influence (i.e., subjects in the Independent treatments), in each task they were asked to follow three steps: (1) review the news, (2) provide a binary judgement on the veracity of the news and report their confidence on the judgement, and (3) indicate the likelihood for them to share the news. In addition, here, if subjects had access to the AI-based credibility indicators (i.e., subject were not in the Control treatment), they would see the AI model's prediction on the news veracity either along with the news in Step 1 (if the subject was in the AI-before treatment), or after Step 2 so that they got a chance to update their veracity judgements and confidence before proceeding on to Step 3 (if the subject was in the AI-after treatment).

4.1.3 Experiment Procedure. Again, we posted our second experiment as a HIT on MTurk to U.S. workers. The procedure of Experiment 2 was largely identical to that of Experiment 1, except for the following differences: (1) Workers who had previously participated in our Experiment 1 were excluded from taking this HIT; (2) upon the approval of a subject, we randomly assigned the subject into one of the six treatments (i.e., {Independent, Non-Independent} × {Control, AI-before, AI-after}); (3) we did *not* tell subjects about the veracity distribution of the news dataset from which the news in the HIT were randomly drawn; (4) for subjects assigned to the the Independent treatments, they followed the 3-step procedure in each task as described in Section 4.1.2 rather than the 5-step procedure.

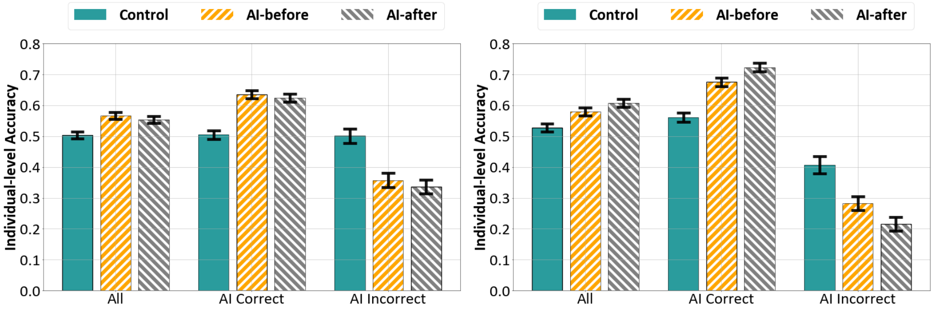
4.1.4 Analysis Methods. To answer **RQ1** and **RQ2**, we restricted our analyses to the data obtained in the Non-Independent treatments, and we adopted the same independent variable, dependent variables, and statistical methods as those used in Experiment 1. In addition, inspired by a few recent research [38, 39], we pre-registered a third dependent variable for **RQ1**—"*truth discernment*", which was calculated as a subject's frequency of labeling a piece of real news as "real" minus the subject's frequency of labeling a piece of fake news as "real" in the final veracity judgement. Different from the two accuracy-related metrics (i.e., individual-level accuracy and sequence-level accuracy), this truth discernment metric captures how much more a subject believed true information *relative to* false information and therefore reflects the subject's sensitivity in distinguishing true and false information. Similarly, for **RQ2**, we also pre-registered an additional dependent variable—"*sharing discernment*", which was a subject's average level of willingness to share real news minus the subject's average level of willingness to share fake news. Intuitively, higher values in truth discernment indicate that people are better at telling apart real and fake news, and higher values in sharing discernment imply a larger decrease in people's engagement with fake news relative to real news.

Finally, to compare the effect size of the AI-based credibility indicators on subjects with and without social influence (i.e., **RQ3**), we adopted the following method: for a dependent variable, we used bootstrapping ($K = 1000$) to re-sample our experimental data and estimated the effect sizes of the AI-before (or AI-after) treatment against the Control treatment as Cohen's d. Such estimation was done within the Independent and Non-Independent treatments separately. We then used paired t-tests to compare the mean value of estimated effect sizes in Non-Independent treatments (i.e., when social influence is present) and the mean value of estimated effect sizes in Independent treatments (i.e., when social influence is absent). We also reported the *probability of superiority* [20], which reflects how often a randomly selected effect size estimate in one group (e.g., with social influence) is larger than a randomly selected effect size estimate in the other group (e.g., without social influence). These analyses were conducted separately for news where the AI model's prediction was correct and incorrect. Moreover, we only conducted these analyses on the individual-level dependent variables as sequence-level dependent variables were not well-defined for subjects who were not exposed to social influence (i.e., subjects in the Independent treatments).

4.2 Results

In total, 1098 workers took our HIT and passed the attention check (35.4% female, with the majority aged between 25 and 44), among whom 476 workers and 622 workers were assigned to the Independent and Non-Independent treatments, respectively. In the following, we highlight our main findings of Experiment 2. For a complete summary of all statistical analyses results, please see the supplemental materials.

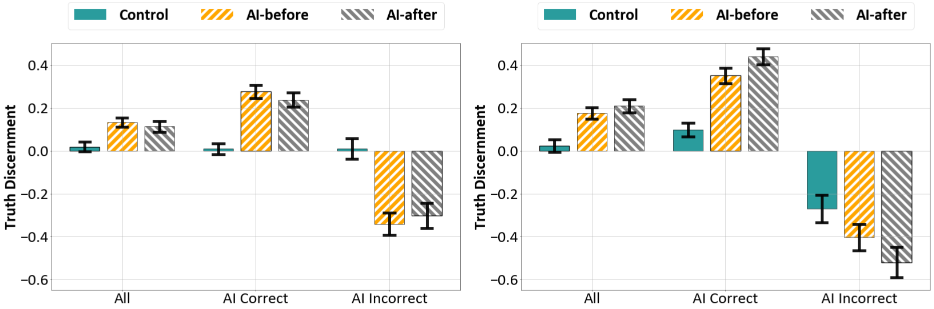
4.2.1 Robustness Check of RQ1. We start by checking whether our answers to **RQ1** (i.e., the effects of AI-based credibility indicators on people's accuracy in detecting misinformation) still hold when subjects were not informed of the distribution of news veracity and the news contained multimedia components. Figure 7a compares the accuracy of subjects' final veracity judgements (i.e., individual-level accuracy) across the three treatments when they were impacted by others' opinions in determining news veracity. Consistent with what we've found in Experiment 1, one-way ANOVA tests suggest significant differences across the three treatments both when the AI model is correct ($F(2, 4099) = 30.23, p < 0.001$) and when the AI model is incorrect ($F(2, 1350) = 15.41, p < 0.001$).



(a) With social influence

(b) No social influence

Fig. 7. The impacts of AI-based credibility indicators on subjects' individual-level accuracy in judging news veracity, for subjects who were subject to social influence (7a) and who were not subject to social influence (7b), respectively. Error bars represent the standard errors of the mean.



(a) With social influence

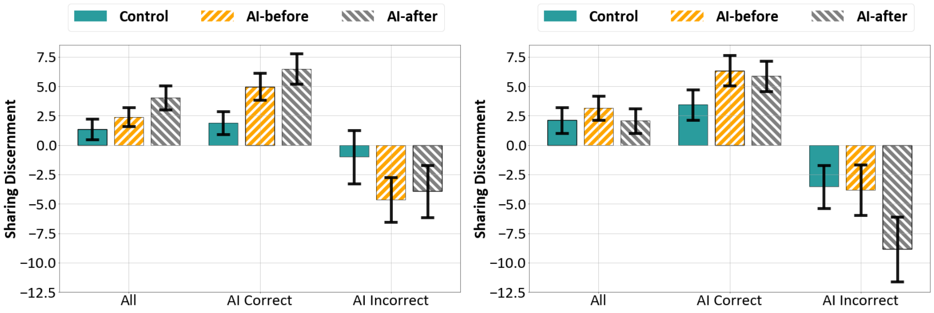
(b) No social influence

Fig. 8. The impacts of AI-based credibility indicators on subjects' truth discernment in judging news veracity, for subjects who were subject to social influence (8a) and who were not subject to social influence (8b), respectively. Error bars represent the standard errors of the mean.

Post-hoc pairwise comparisons further show that the significant differences are caused by the presence of the AI-based credibility indicators rather than their timing. When we examine the impact of AI-based credibility indicators on the sequence-level accuracy, we again get similar conclusions as those in Experiment 1—the crowd's accuracy in judging the news veracity significantly increases when a correct AI prediction is provided and decreases when an incorrect AI prediction is provided (see supplementary materials for additional figures).

In addition, Figure 8a shows the comparison of individual subject's truth discernment across the three treatments. Again, we find that the differences are significant both when the AI model is correct ($F(2, 565) = 23.06, p < 0.001$) and when the AI model is wrong ($F(2, 316) = 13.45, p < 0.001$), while the timing of when the AI-based credibility indicators are shown doesn't have a clear impact on subjects' truth discernment. In other words, when people's interpretation of a piece of news is influenced by others' judgements on its veracity, providing a correct (incorrect) AI-based credibility indicator significantly increases (decreases) people's capability to differentiate real news and fake news, though whether the credibility indicator is provided before or after people form their independent opinions doesn't seem to matter.

Together, our results here suggest that under the new and more realistic settings of Experiment 2, our answers to **RQ1** still hold.



(a) With social influence

(b) No social influence

Fig. 9. The impacts of AI-based credibility indicators on subjects' sharing discernment, for subjects who were subject to social influence (9a) and who were not subject social influence (9b), respectively. Error bars represent the standard errors of the mean.

4.2.2 Robustness Check of RQ2. Repeating the similar analyses as what we've done in Experiment 1, we again find consistent results with respect to the effects of AI-based credibility indicators on people's willingness to share real and fake news when social influence is present (i.e., **RQ2**). Specifically, when the AI model provided a correct prediction on the news veracity, subjects tended to report slightly higher levels of willingness to share real news (AI-before − Control: $\Delta M = 2.25$, AI-after − Control: $\Delta M = 0.78$) and slightly lower levels of willingness to share fake news (AI-before − Control: $\Delta M = -1.52$, AI-after − Control: $\Delta M = -3.78$), though these differences are not statistically significant ($p > 0.05$). Our statistical tests on the expected depth of spread also suggest that the differences between the AI-before/AI-after treatments and the Control treatment are not statistically significant ($p > 0.05$), both for real news and fake news. Finally, we obtain similar observations in those cases when the AI model provides an incorrect prediction on the news veracity—the provision of incorrect AI-based credibility indicators does not exhibit a significant impact on either the individual's sharing intention or the expected depth of spread for a piece of news, regardless whether the news is real or fake.

Interestingly, as shown in Figure 9a, we detect some differences across the three treatments in terms of individual subject's sharing discernment. One-way ANOVA tests suggest that when the AI model's veracity prediction on a piece of news is correct, the differences in sharing discernment are statistically significant across the three treatments ($F(2, 565) = 4.04, p = 0.018$), and post-hoc pairwise comparisons indicate that the significant increase in sharing discernment is *only* observed between the Control and AI-after treatment ($p = 0.015$). This implies that providing correct AI model predictions on news veracity only nudges people into decreasing their sharing of fake news relative to real news, when these predictions are shown to people after they have formed their independent judgements about the news. On the other hand, while we also see a similar trend in Figure 9a that the provision of incorrect AI-based credibility indicators leads to a decreased level of sharing discernment, this decrease is not significant according to our one-way ANOVA test.

4.2.3 Compare effect sizes when social influence is present/absent. Finally, we look into **RQ3** to compare the effect sizes of the AI-based credibility indicators in influencing people's detection and spread of misinformation, between the case when social influence is present and the case when social influence is absent. To provide some visual intuition first, we include in Figures 7b, 8b, and 9b the illustrations of the effects of AI-based credibility indicators on subjects' individual-level accuracy, truth discernment, and sharing discernment, respectively, when social influence is *absent*. One can directly compare the observed effects in these figures with the corresponding ones in

AI Correctness	Dependent Var	d (Independent)	d (Non-Independent)	$\Delta\bar{d}$	Prob. of Superiority
AI Correct	Individual accuracy	0.24 [0.16, 0.32]	0.27 [0.19, 0.34]	0.03^{***}	0.68
	Truth discernment	0.61 [0.38, 0.84]	0.70 [0.50, 0.90]	0.09^{***}	0.72
	Sharing discernment	0.18 [-0.05, 0.39]	0.21 [-0.01, 0.41]	0.02^{***}	0.58
AI Incorrect	Individual accuracy	0.27 [0.12, 0.41]	0.30 [0.16, 0.44]	0.03^{***}	0.59
	Truth discernment	0.25 [-0.07, 0.55]	0.69 [0.40, 0.98]	0.44^{***}	0.98
	Sharing discernment	0.02 [-0.29, 0.32]	0.17 [-0.10, 0.43]	0.14^{***}	0.76

Table 1. Comparison of effect sizes of the AI-BEFORE treatment. d (Independent) and d (Non-Independent) report the treatment's effect sizes (in terms of Cohen's d) and the 95% bootstrap confidence intervals when social influence is absent and present, respectively. $\Delta\bar{d}$ is the difference of the average effect sizes; a positive value suggests that the effect size is larger when social influence is present. Paired t-tests are used to examine whether the differences in effect sizes are statistically significant, and results are reported in superscripts along with $\Delta\bar{d}$, with *** representing a significance level of 0.001. Probability of superiority reports the chance that a randomly selected effect size estimate from the case when social influence is present is larger than that from the case when social influence is absent; the larger the probability the more certain we are that in any random experiment, the effect size is larger when social influence is present.

AI Correctness	Dependent Var	d (Independent)	d (Non-Independent)	$\Delta\bar{d}$	Prob. of Superiority
AI Correct	Individual accuracy	0.34 [0.25, 0.43]	0.24 [0.16, 0.32]	-0.10^{***}	0.05
	Truth discernment	0.84 [0.58, 1.10]	0.57 [0.37, 0.77]	-0.26^{***}	0.05
	Sharing discernment	0.16 [-0.08, 0.39]	0.29 [0.10, 0.49]	0.13^{***}	0.80
AI Incorrect	Individual accuracy	0.42 [0.26, 0.59]	0.33 [0.20, 0.47]	-0.09^{***}	0.21
	Truth discernment	0.44 [0.08, 0.80]	0.57 [0.30, 0.88]	0.13^{***}	0.72
	Sharing discernment	0.26 [-0.10, 0.58]	0.13 [-0.14, 0.39]	-0.13^{***}	0.28

Table 2. Comparison of effect sizes of the AI-AFTER treatment. d (Independent) and d (Non-Independent) report the treatment's effect sizes (in terms of Cohen's d) and the 95% bootstrap confidence intervals when social influence is absent and present, respectively. $\Delta\bar{d}$ is the difference of the average effect sizes; a positive (negative) value suggests that the effect size is larger (smaller) when social influence is present. Paired t-tests are used to examine whether the differences in effect sizes are statistically significant, and results are reported in superscripts along with $\Delta\bar{d}$, with *** representing a significance level of 0.001. Probability of superiority reports the chance that a randomly selected effect size estimate from the case when social influence is present is larger than that from the case when social influence is absent; the larger the probability the more certain we are that in any random experiment, the effect size is larger when social influence is present.

Figures 7a, 8a, and 9a to infer how the sizes of these effects compare to those in the case when social influence is *present*.

To conduct this comparison formally, we follow the method that we've described in Section 4.1.4 to generate bootstrapped samples of our experimental data and estimate the effect sizes of a treatment—both with and without social influence—as Cohen's d based on the bootstrapped samples. We report the effect size estimation results in Tables 1 and 2, for the AI-BEFORE and AI-AFTER treatments, respectively (see the "d(Independent)" and "d(Non-Independent)" columns)[9]. Note that since we were not able to find any significant impacts in terms of the presence and timing of AI-based credibility indicators on subject's willingness to share real news or fake news (see Section 4.2.2), we omit the effect size comparison on individual's sharing intention.

[9] When the AI model's prediction is correct, the CONTROL treatment is treated as the reference; when the AI model's prediction is incorrect, the AI-BEFORE or AI-AFTER treatment is treated as the reference. Doing so, all estimated effect sizes take positive values and are easier to interpret.

As shown in Table 1, if the AI-based credibility indicators are provided along with the news *before* people form their independent opinions about the news, it appears that their effects on people's detection and spread of misinformation are consistently *larger* when social influence is present, compared to when social influence is absent. Indeed, our paired t-tests results confirm that the *average* effect size of the AI-BEFORE treatment is significantly larger ($p < 0.001$) when subjects were exposed to social influence, regardless of the correctness of the AI model and the dependent variable we consider (see "$\Delta\overline{d}$" column in Table 1). Moreover, to inform that *in any random run of the experiment*, how likely the effect size of the AI-BEFORE treatment is larger when social influence exists (as opposed to whether *on average*, the effect size of the AI-BEFORE treatment is larger when social influence exists), we further report the probability of superiority in Table 1. As these probabilities are consistently larger than 0.5, we are reasonably confident about our conclusion that the effect size of the AI-BEFORE treatment is larger when social influence is present. Interestingly, a closer look at Figures 7–9 indicates that when social influence is present, subject's individual-level accuracy, truth discernment, and sharing discernment do not seem to reach a level that is significantly different from those in the case when social influence is absent; so the increase of effect size is mainly driven by the differences in the CONTROL treatments. For example, on the news where the AI model's prediction is correct, subjects in the NON-INDEPENDENT, CONTROL treatment tended to have lower individual-level accuracy, truth discernment, and sharing discernment than subjects in the INDEPENDENT, CONTROL treatment, possibly as they were influenced by others' incorrect veracity judgements; thus, the effect sizes of the AI-BEFORE treatment become larger with social influence potentially because providing the correct AI-based credibility indicators before people form their independent opinions cancels out or at least reduces the negative impacts brought up by others' incorrect judgements.

In contrast, the comparison of the effect sizes for the AI-AFTER treatment between the cases with or without social influences is less clear. As shown in Table 2, we find that in most cases, especially with respect to subjects' individual-level accuracy in detecting misinformation, the impacts of showing the AI-based credibility indicators *after* people have formed their independent opinions seem to be *weakened* by the social influence. However, we also note that when social influence exists, the effect size of the AI-AFTER treatment seems to be larger on sharing discernment when the AI model's prediction is correct, and on truth discernment when the AI model's prediction is incorrect, although Figures 8–9 again suggest that this observation is mainly driven by the differences across the two CONTROL treatments.

5 DISCUSSIONS

In this section, we first summarize the potential benefits and risks of AI-based credibility indicators as well as the limitations of these indicators. We then provide implications on better utilizing AI to combat misinformation, as well as designing experimental research to study the effects of news credibility indicators in more realistic settings. Finally, we discuss the limitations of our study.

5.1 Benefits, Risks, and Limitations of AI-based Credibility Indicators

The results of our study suggest that leveraging AI-based credibility indicators to help people identify misinformation comes with benefits, risks, as well as limitations, when these people are subject to social influence in judging the credibility of online information. On the positive side, our study shows that even if people are influenced by others when judging the veracity of news, providing *accurate* AI-based credibility indicators along with news items can effectively improve people's ability in detecting misinformation. This highlights the promise of utilizing automated AI technologies to significantly speed up the evaluation of the quality of online information. Indeed, the traditional approach of recruiting fact-checking professionals to manually check the reliability of each news is not only expensive, but also falls short in catching up with the unprecedented speed that information is generated and spread today. Even more worrisome, the sparse application of warning labels on news stories due to the limited scalability of manual fact-checking can even bring up an unintended side effect called the "implied truth effect" [35], that is, people may consider false information without warning labels to be true as they incorrectly assume these information have already been verified. In light of this, reliable AI-based misinformation detection technologies can be especially beneficial as they can be used to signal the credibility of *all* of the online information almost in real-time, while they still exhibit a high degree of effectiveness in influencing people's perceptions of the information.

However, our study results also reveal that people, even together with their peers, lack the capability in determining the correctness of AI-based credibility indicators. This implies serious risks of people being misled by AI-based credibility indicators that are wrong, and consequently believing in or even spreading misinformation. In our exploratory analysis (Section 3.5.3), we show that the presence of AI-based credibility indicators significantly increases people's tendency to align their final veracity judgement on a piece of news with the model's prediction when their initial judgement disagrees with the model. In fact, if we restrict our attention to only those cases where the AI-based credibility indicator is wrong while an individual's initial veracity judgement on the news is correct, we still find that the individuals will significantly increase the likelihood of switching their belief in the veracity of the news to the wrong prediction despite their independent predictions being correct, when the incorrect AI-based credibility indicator is shown to them as a reference together with others' veracity judgements on the news. This is true even when both the individuals' independent veracity judgements and the majority of other people's veracity judgements on the news are correct. Together, these results imply the urgent need of assisting people to effectively gauge the accuracy of AI-based credibility indicators, so that they can make use of these indicators more appropriately.

Finally, we note that when people are exposed to social influence in evaluating the veracity of news stories, it appears that AI-based credibility indicators can not unlock their full potential in influencing people's perception of and engagement with the news. As discussed in Section 3.5.3, the presence of AI-based credibility indicators is not able to "strengthen" an individual's veracity belief in news when the individual's initial veracity judgement aligns with the AI model's prediction. We find this is still true when we focus on only those cases where both the individual's initial veracity judgement and the AI model's prediction are correct. In other words, even if the AI-based credibility indicators are perfectly reliable, the agreement between the AI model and one's own judgement on the veracity of a piece of news does *not* motivate people to stick with their correct judgement to a higher extent, regardless of whether the majority of other people agrees with the AI model or not. This phenomenon may have partly contributed to our observation that the positive impacts brought up by accurate AI-based credibility indicators on the crowd's accuracy in detecting misinformation are not amplified by social influence as the news spreads into a larger number of people. It is thus an important future work to explore how to further release the potential of AI-based credibility indicators to strengthen people's belief in their correct judgements via the agreement between people and the AI model.

5.2 Implications for Better Utilizing AI to Combat Misinformation

The potential risks of supplying AI-based credibility indicators, as we have discussed in Section 5.1, suggest AI technologies should be incorporated into the fight against misinformation with extra care. One possible approach is to adopt a hybrid, human-AI collaborative fact-checking procedure, which may be implemented either through a machine-in-the-loop paradigm where fact-checkers produce all the final credibility warning labels but get recommendations from AI models [34], or a human-in-the-loop paradigm where the AI models actively decide when they need human inputs [45]. In the former case (i.e., machine-in-the-loop), one critical challenge to address is to design the AI models to optimize for the human-AI *joint* decision-making outcome [8]. Addressing this challenge requires a better understanding on both how to tune the AI model to complement human fact-checkers, and how to structure the information exchange between human fact-checkers and the AI model to enhance *fact-checkers*' trust calibration in the AI (e.g., via presenting model explanation and confidence [53, 57], allowing fact-checkers to interact with the model [34], etc.). In the latter case (i.e., human-in-the-loop), beyond a similar challenge of optimizing the design of AI-based credibility indicators so that *end-users* can effectively determine the reliability of these indicators, a few additional interesting questions to ask include whether and how the AI model can solicit the wisdom of a mixed-expertise crowd (e.g., professional fact-checkers, domain experts, and a community of laypeople), as well as whether and how to present credibility indicators generated from various channels (e.g., AI, AI+fact-checkers, AI+community, AI+community+fact-checkers) to end-users to maximize their desirable effects.

5.3 Towards Studying Credibility Indicators in More Realistic Settings

The fight against misinformation is of great societal importance, which urges researchers to maximize the ecological validity of their research, especially for those related to understanding the effectiveness of interventions. To this end, conducting field experiments is a straight-forward solution (e.g., [32, 37]) whenever possible. We believe another approach for increasing the ecological validity of misinformation-related research is to design more realistic experimental settings in controlled experiments. In this study, we make the first attempt to study the effects of AI-based credibility indicators in a more realistic setting by taking social influence into consideration. Our results provide us with useful understandings that we would not have been able to get if the experiment is designed in a simplified setting, such as the size of AI-based credibility indicators' impacts on people in a real social environment may be larger than what we would conclude from a controlled experiment with no social influence, when these indicators are presented to people before they form their own judgements about the news. We acknowledge that, however, our experimental setting is still far away from what a natural social media environment looks like, and there are plenty of opportunities for creating more realistic experimental settings to study the effects of credibility indicators in future work. For example, previous research has shown that people suffer from the "illusory truth effect" when judging the credibility of online information, which suggests that they tend to consider repeatedly seen and thus familiar information to be more true than novel information [36]. Taking the impacts of both social influence and repeated exposure on one's perceptions of the news into consideration requires us to place human subjects into a full-fledged social network rather than a diffusion "path" in the network. As another example, in our study, the social influence that subjects get exposed to when evaluating news veracity essentially comes from a random population. However, in the real world, people have the tendency to befriend with like-minded individuals, which contributes to the creation of filter bubbles [7] and may affect the extent to which the views of those people that one connects to are polarized. Thoroughly understanding the effects of AI-based credibility indicators in the presence of selective exposure is therefore another interesting future work.

5.4 Limitations

Our study was conducted with laypeople (i.e., subjects recruited from Amazon Mechanical Turk) on one specific type of news (i.e., news related to COVID-19). Cautions should be used when generalizing results in this work to different settings, such as when there exist some domain experts in the crowd, or when the news is about a different topic. In particular, Horne et al. [23] find that people are more receptive to advice from AI when evaluating the veracity of news topics that are novel and evolving, and we consider COVID-19 as such a topic. Thus, more experimental studies should be carried out in the future with recurring news topics (e.g., climate change) to understand to what extent the results reported here can be generalized when people have strong prior beliefs on the topic. Another limitation of our study is that our experimental setup can not reflect the "closeness" between people who review the same news. In reality, individuals' perceptions of a piece of news may be influenced by those people who they consider as close to them to a larger extent; thus, it is unclear whether our experimental results will generalize to the scenario when different people influence one's judgements to a different degree. We also note that the format of the news stories that we presented to our subjects was fairly simple—e.g., only the text/image of the news stories was shown without any sources of the news being displayed. As earlier research shows that information such as the source of news stories may serve as a critical heuristic for people to gauge its credibility [23, 39], it would be interesting to explore in the future how contextual information (e.g., news sources) and social influence, together, impact people's perceptions of and engagement with the news.

6 CONCLUSIONS

In this paper, we present two randomized human-subject experiments to understand the effects of AI-based credibility indicators on people while taking social influence into consideration. We find that even when people's perceptions of the news are influenced by others' opinions about it, presenting the AI-based credibility indicators along with the news can still nudge people into aligning their veracity belief in the news with the AI model's prediction, regardless of the correctness of the prediction, thereby changing people's ability in detecting misinformation. However, these AI-based credibility indicators exhibit limited impacts on people's engagement with the news under the social influence, regardless of whether it is real or fake. We also find that compared to the case when social influence is absent, providing AI-based credibility indicators to people before they make their independent judgements of the news results in a larger impact on people's perception of and engagement with the news when people are subject to social influence. Our results provide important insights into understanding the possible benefits, risks, and limitations of AI-based credibility indicators, and we discuss practical implications on better utilizing AI technologies to combat misinformation and on improving the design of controlled experiments to study misinformation in more realistic settings.

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-1850335, and the Purdue Research Foundation under Ross-Lynn Research Scholar Fund on this work. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors alone.

REFERENCES

[1] Jonathan E Alevy, Michael S Haigh, and John A List. 2007. Information cascades: Evidence from a field experiment with financial market professionals. *The Journal of Finance* 62, 1 (2007), 151–180.

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

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

[4] Lisa R Anderson and Charles A Holt. 1997. Information cascades in the laboratory. *The American economic review* (1997), 847–862.

[5] Mihai Avram, Nicholas Micallef, Sameer Patil, and Filippo Menczer. 2020. Exposure to social engagement metrics increases vulnerability to misinformation. *arXiv preprint arXiv:2005.04682* (2020).

[6] Joseph B Bak-Coleman, Ian Kennedy, Morgan Wack, Andrew Beers, Joseph S Schafer, Emma S Spiro, Kate Starbird, and Jevin D West. 2022. Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour* (2022), 1–9.

[7] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.

[8] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. 2021. Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11405–11414.

[9] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[10] 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.

[11] Nadia M Brashier, Gordon Pennycook, Adam J Berinsky, and David G Rand. 2021. Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences* 118, 5 (2021).

[12] Michele Cantarella, Nicolò Fraccaroli, and Roberto Volpe. 2020. Does fake news affect voting behaviour? (2020).

[13] Vincenzo Carrieri, Leonardo Madio, and Francesco Principe. 2019. Vaccine hesitancy and (fake) news: Quasi-experimental evidence from Italy. *Health economics* 28, 11 (2019), 1377–1382.

[14] Chun-Wei Chiang and Ming Yin. 2021. You'd better stop! Understanding human reliance on machine learning models under covariate shift. In *13th ACM Web Science Conference 2021*. 120–129.

[15] Kon Shing Kenneth Chung, Liaquat Hossain, and Joseph Davis. 2007. Individual performance in knowledge intensive work through social networks. In *Proceedings of the 2007 ACM SIGMIS CPR conference on Computer personnel research: The global information technology workforce*. 159–167.

[16] 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.

[17] 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.

[18] Limeng Cui and Dongwon Lee. 2020. CoAID: COVID-19 Healthcare Misinformation Dataset. arXiv:2006.00885 [cs.SI]

[19] 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.

[20] William P Dunlap. 1994. Generalizing the common language effect size indicator to bivariate normal correlations. *Psychological Bulletin* 116, 3 (1994), 509.

[21] Jacob K Goeree, Thomas R Palfrey, Brian W Rogers, and Richard D McKelvey. 2007. Self-correcting information cascades. *The Review of Economic Studies* 74, 3 (2007), 733–762.

[22] 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.

[23] 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.

[24] Elle Hunt. 2017. 'Disputed by multiple fact-checkers': Facebook rolls out new alert to combat fake news. https://www.theguardian.com/technology/2017/mar/22/facebook-fact-checking-tool-fake-news.

[25] 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.

[26] Travis Kriplean, Caitlin Bonnar, Alan Borning, Bo Kinney, and Brian Gill. 2014. Integrating on-demand fact-checking with public dialogue. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1188–1199.

[27] Vivian Lai, Han Liu, and Chenhao Tan. 2020. " Why is' Chicago'deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] Mohsen Mosleh, Cameron Martel, Dean Eckles, and David Rand. 2021. Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[33] Mohsen Mosleh, Gordon Pennycook, and David G Rand. 2020. Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter. *Plos one* 15, 2 (2020), e0228882.

[34] 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.

[35] 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.

[36] Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general* 147, 12 (2018), 1865.

[37] 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.

[38] 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.

[39] Gordon Pennycook and David G Rand. 2021. The psychology of fake news. *Trends in cognitive sciences* (2021).

[40] 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.

[41] Amy Rechkemmer and Ming Yin. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems*. 1–14.

[42] 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.

[43] Yoel Roth and Nick Pickles. 2020. Updating our approach to misleading information. https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.

[44] 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.

[45] Shaban Shabani and Maria Sokhn. 2018. Hybrid machine-crowd approach for fake news detection. In *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*. IEEE, 299–306.

[46] Petter Törnberg. 2018. Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS one* 13, 9 (2018), e0203958.

[47] Brett Trueman. 1994. Analyst forecasts and herding behavior. *The review of financial studies* 7, 1 (1994), 97–124.

[48] Sho Tsugawa and Sumaru Niida. 2019. The impact of social network structure on the growth and survival of online communities. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 1112–1119.

[49] John C Turner. 1991. *Social influence.* Thomson Brooks/Cole Publishing Co.

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

[51] 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.

[52] E Walther and Hartmut Blank. 2004. Entscheidungsprozesse im Falschinformationsparadigma: Die Rolle von Unsicherheit, Metakognition und sozialem Einfluss= Decision processes in the misinformation paradigm: The role of uncertainty, metacognition, and social influence. *Psychologische Rundschau* 55, 2 (2004), 72–81.

[53] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*. 318–328.

[54] 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.

[55] 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.

[56] 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.

[57] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.

Received January 2022; revised April 2022; accepted May 2022