Trust It or Not: Effects of Machine-Learning Warnings in Helping Individuals Mitigate Misinformation

Haeseung Seo
hxs378@psu.edu
The Pennsylvania State University
University Park, PA, USA

Aiping Xiong
axx29@psu.edu
The Pennsylvania State University
University Park, PA, USA

Dongwon Lee
dongwon@psu.edu
The Pennsylvania State University
University Park, PA, USA

ABSTRACT

Despite increased interests in the study of fake news, how to aid users' decision in handling suspicious or false information has not been well understood. To obtain a better understanding on the impact of warnings on individuals' fake news decisions, we conducted two online experiments, evaluating the effect of three warnings (i.e., one Fact-Checking and two Machine-Learning based) against a control condition, respectively. Each experiment consisted of three phases examining participants' recognition, detection, and sharing of fake news, respectively. In Experiment 1, relative to the control condition, participants' detection of both fake and real news was better when the Fact-Checking warning but not the two Machine-Learning warnings were presented with fake news. Post-session questionnaire results revealed that participants showed more trust for the Fact-Checking warning. In Experiment 2, we proposed a Machine-Learning-Graph warning that contains the detailed results of machine-learning based detection and removed the source within each news headline to test its impact on individuals' fake news detection with warnings. We did not replicate the effect of the Fact-Checking warning obtained in Experiment 1, but the Machine-Learning-Graph warning increased participants' sensitivity in differentiating fake news from real ones. Although the best performance was obtained with the Machine-Learning-Graph warning, participants trusted it less than the Fact-Checking warning. Therefore, our study results indicate that a transparent machine learning warning is critical to improving individuals' fake news detection but not necessarily increase their trust on the model.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**;

KEYWORDS

Misinformation; warnings; trust; algorithm transparency

ACM Reference format:
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 Websci '19: 11th ACM Conference on Web Science, Boston, MA, June 30–July 03, 2019 (Websci '19),* 10 pages.
https://doi.org/10.1145/1122445.1122456

Websci '19, June 30–July 03, 2019, Boston, MA

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

We currently live in a historical era called the "information age". The advent of modern information technology fundamentally changes the ways how people access, communicate and share information. Specifically, the rise of the Internet and more recently social media platforms (e.g., Facebook, Twitter) have made it possible for individuals to produce, consume, and share diverse multi-modal information (e.g., text, picture, video). With the boundary between information source and information receiver becoming blurred and often invisible, then, issues arise with regard to the quantity and quality of the information to which people are exposed [3]. Especially, it must be acknowledged that people are not necessarily good at evaluating the quality of online information. *Fake News* often refers to (intentionally) false stories or fabricated information written and published for various incentives including political agenda or financial gain [8, 13, 34]. In recent years, the spread of fake news has been identified as a major risk for individuals and society [35]. For instance, fake news has fostered people's bias and false belief of climate change [39] and greatly influenced elections and democracies [6].

Two venues of approaches have been investigated to mitigate the negative impacts of fake news: (1) computation-based detection and prevention of fake news; (2) decision-aid methods to warn users when a piece of fake news has been identified. Among the latter venue of approaches (the focus of this study), attaching warnings to the news that was suspicious or fact-checked to be fake news was implemented to discourage users' consumption and belief in fake news. One such example once was used by Facebook. While some studies showed that exposure to a fact-checking warning under Facebook style headlines reduced the perceived accuracy of fake news compared to a control condition [9], other studies did not [27], motivating our study.

Also, with more fact-checking work being done by machine learning algorithms [35], one interesting but rarely investigated question related to both venues is: *After computational methods detect fake news, how to convincingly present the result to users to make informed decisions consequently?* To answer this intriguing question, we investigate the following research questions:

(1) RQ1: Will the presence of a fact-checking warning increase participants' fake news detection relative to a control condition in which there is no warning?

(2) RQ2: Will the presence of automatic fake news detection results using machine learning algorithms increase participants' fake news detection relative to the control condition?
(3) RQ3: What is the best way to communicate the result of machine learning based on fake news detection?

In our study, we proposed new machine-learning warnings in response to an emphasis on "algorithm transparency" [11, 26, 31]. A *Fact-Checking* warning that was used in the study of [9] was also used to see whether we could replicate their results. Using a between-subjects design, we conducted two online experiments on Amazon Mechanical Turk (MTurk), in each of which the immediate, short-term, and long-term effectiveness of three warnings against a control condition was evaluated in three phases, respectively. Across two experiments, 1,176 MTurk workers completed three interrelated decision tasks of *recognition*, *detection*, and *sharing* to different news (half real and half fake) in each phase. In addition to the analysis of decision rates, we used a *signal-detection theory* (SDT) [24, 37] approach assessing individuals' susceptibility and bias at detecting fake news.

Across all phases of Experiment 1, participants showed limited recognition and cautious sharing decisions in general. Compared to the control condition, participants increased their correct detection of both fake and real news in the *Fact-Checking* condition but not the others. In Experiment 2, when the news source, a cue that most participants used to identify news' legitimacy, was removed from each news headline, similar results were obtained for the recognition and sharing tasks. But the effect of the *Fact-Checking* warning obtained in Experiment 1 disappeared. Instead, compared to the control condition, a *Machine-Learning-Graph* warning increased participants' sensitivity in differentiating fake and real news.

Our work makes the following three key contributions:

(1) We proposed and evaluated the use of warnings to communicate the results of machine-learning detected fake news to users. Across three machine-learning warnings, only the *Machine-Learning-Graph* warning that includes the detail results of machine-learning based detection increased individuals' correct detection of fake news, suggesting that a transparent machine learning algorithm is critical to improve people's fake news detection.
(2) Our results showed that the *Fact-Checking* warning increased participants' correct detection of both fake and real news when the source was included in news headlines but not when the source was excluded. Participants showed more trust on the *Fact-Checking* warning even though the best detection performance was obtained with the *Machine-Learning-Graph* warning, suggesting promoting users' fake news detection does not necessarily promoting users' trust on the warning.
(3) We introduced a SDT approach to investigate individuals' fake news detection and obtained that the *Machine-Learning-Graph* warning increased participants' sensitivity to differentiate fake from real news but not the *Fact-Checking* warning.

These contributions bridge the two venues of fake news mitigation and should help researchers and practitioners improve their understanding of people's decision-making in facing fake news and propose usable and transparent algorithms to address fake news problems.

2 RELATED WORK

2.1 Human Fake News Detection

Within experimental settings, a few factors have been investigated to understand their impact on people's belief in and willingness to share fake news on social media. Pennycook et al. [27] conducted online studies examining the influence of warning and repetition. In their Experiments 2 and 3, participants were asked to evaluate different pieces of news in multiple stages. In stage 1, participants were asked to indicate whether they were to share news headlines (half fake and half real) on social media. Also, half of the participants were randomly assigned to a warning condition, in which all fake news stories were flagged with a caution symbol and the text "Disputed by 3rd Party Fact-Checkers". The rest half were assigned to a control condition in which no warning was presented. After a distracting stage, in stage 3, participants were asked to rate familiarity and accuracy of real and fake news headlines (a half from stage 1 and a half from a new set of headlines). Each participant in Experiment 3 was also invited to return for a follow-up session one week later in which the same headlines were seen in stage 3 and a new set of headlines were presented. Results showed that repeated headlines were rated as more "real" than novel headlines regardless of headlines' legitimacy and warning. The increased accuracy perception obtained with a single exposure lasted even after a week regardless of the warning. Although the main effect of warning and its interaction with news legitimacy were significant in Experiment 2, neither terms were significant in Experiment 3.

Clayton et al. [9] conducted an online study to further investigate the effect of warning. To eliminate confounding variables, they removed the source within all news headlines. In one condition, they implemented a "Fact-Checking" warning similar to that in [27] but specified the third parties' names within the warning. 413 participants in the condition indicated their perceived accuracy and likelihood to "Like" or share nine news (six fake, four of which with a warning, and three real). Compared to a control condition, participants' perceived accuracy of fake news with the warning was reduced, indicating the effectiveness of using warning to reduce participants' belief in fake news.

A comparison between the studies of [27] and [9] revealed several critical differences, which may cause the ineffectiveness of the warning in Experiment 3 of [27] but the effect obtained by [9]. First, warnings were presented at the familiarity phase of [27] but the evaluation phase of [9]. Thus, Clayton et al. [9] evaluated the effect of warning but Pennycook et al. [27] evaluated its short-term effect. Second, the source was removed for each news headline used by [9], which may increase participants' reliance on using warning to assess the legitimacy of news headlines. Also, the 3rd party names were specified in [9], which may increase individuals' trust on the warning. Accordingly, in our work, we investigated a warning like [9] during the assessment phase but varied the presence and absence of the source to understand how it impacts individuals' belief in fake news with warnings. Besides, we evaluated the immediate, short-term, and long-term effects of the warning in different phases, and asked participants to indicate their trust level on the warning.

2.2 Computational Fake News Detection

In recent years, much attention has been made to detect fake news using computational means (e.g., [10, 36, 38]), especially using various features such as single-modal [4, 30] and multi-modal features [20, 40]. The single-modal methods mainly focus on analyzing the textual contents of news, for example, counting the number of assertive words which are shown more in trusted sources [30] or evaluating the consistency between topic sentence and main text [4]. Meanwhile, multi-modal methods include features derived from various sources, such as contents of news, users who posted news, publishers of news, or how news has propagated in a network. For instance, those features can be several textual features including news contents and user's comments [33] or different data types including a combination of text, image, or video [20, 25, 40].

In addition, to provide the accountability of algorithmic solutions, researchers have started offering details about the inner mechanisms of machine learning algorithms [5]. With more fact-checks done by machine learning algorithms [35], we study how to present the result *after* the detection of fake news occurs. Specifically, the machine-learning warnings in our study were not generated by machine learning algorithms. Instead, we used hypothetical evaluation metrics (e.g., accuracy) and multi-modal features (e.g., text, picture) of machine learning algorithms within various warning signs (e.g., one with the wording "Machine Learning") to leverage the advancements in computational solutions.

2.3 Signal Detection Theory

Accuracy measure, such as the number of correct identification of fake news, is incomplete to understand individuals' vulnerability to fake news because they ignore factors, such as the influence of real news. Accordingly, in our work, in addition to measures of decision rates, we use SDT [16] to understand individuals' detection in response to fake news. SDT has been implemented for investigating decision-making in the context of perceptual uncertainties and risk [24], such as susceptibility to a phishing email and web pages [7, 42].

In SDT, participants' responses are defined as two normal distributions of pieces of evidence, representing both *signal* and *noise*. The difference between the means of signal and noise distributions reflects participants' sensitivity (d'), e.g., their ability to tell whether a piece of news is fake. Independent from d', SDT also allows a measure of participants' response criterion (c), e.g., their tendency to treat a piece of news as fake. In the context of fake news detection, the signal will be fake news to detect and the noise will be real news. If the news is fake and the decision for the news judgment is suspicious, the trial is a *H*: hit. If there is a piece of real news but is judged suspicious, it is a *FA*: false alarm. If fake news is misjudged as non-suspicious, it is a miss. Finally, if real news is judged as non-suspicious, it is a correct decision. d' and c are derived as follows:

$$d' = z(H) - z(FA) \quad (1)$$

$$c = -0.5[z(H) + z(FA)] \quad (2)$$

Therefore, using SDT, the evaluation of how well a participant detects fake news will be not influenced by whether the participant is biased or not.

2.4 Current Work

The question of whether machine learning warning reduces individuals' fake news susceptibility has consequences for a broader perspective on the deployment of transparent machine learning algorithms. In this paper, we conducted two experiments investigating the three *RQs* by examining participants' recognition, detection, and willingness to share fake and real news. The detailed data from all our experiments is available for download at http://pike.psu.edu/download/websci19/.

3 EXPERIMENT 1

We conducted a between-subjects online study investigating the effect of two machine-learning and one fact-checking warning in mitigating fake news. In addition to the three warning conditions, a control group (*CON*) in which no warning was presented, was also included in the study. Participants made recognition, detection, and sharing decisions on fake and real news in three phases. In Phase 1, participants got warnings on fake news trials except for those participants in *CON*. After a distraction task of filling demographic information, Phase 2 started, in which participants did the same task as Phase 1 without warning to evaluate the short-term effect of the warning. One week later, we invited each participant back to Phase 3 to do the same task as Phase 2 to evaluate the long-term effect of the warning. Half of the trials in Phase 3 were news headlines that were already presented in Phases 1 and 2, which were used to investigate participants' decisions of repeated fake news.

3.1 Methodology

The study was conducted on Amazon MTurk, and all participants were (1) at least 18 years old; (2) located at the United States; and (3) with a human intelligence task approval rate above 95%. Participants were allowed to participate in the study once. Our online study was programmed using Qualtrics. This and the following study were approved by the Institutional Research Board of The Pennsylvania State University.

Materials. We created 24 news headlines in the format of Facebook posts, consisting of a picture, source, header, and a short description (see Figure 1). 12 were verified fake news from *snope.com* and *politifact.com*, well-known third-party fact-checking websites. The other 12 news headlines were real news chosen from major news media, such as *huffpost.com* and *reuters.com*. The 24 pieces of news were divided into three groups (half real and half fake in each group). For each condition, a Latin-square design was implemented to balance the order of the groups across three phases. We proposed three warnings: Fact-Checking (*FC*), Machine-Learning (*ML*), and Machine-Learning-Accuracy (*MLA*). Each warning was attached to the bottom of the fake news in the study. Figure 1 gives a depiction of the warning design and the content of each warning. The two machine-learning warnings were the same except a hypothetical value, 97%, was described in the *MLA* warning to indicate the accuracy of the machine learning algorithm.

The selected news was released from April to June in 2018, and the topic of news was limited to politics because 1) political news is one type of the most popular news that most individuals will read every day, so most of the people have a certain sense to judge

Figure 1: Warnings presented in Experiment 1, top row: A piece of fake news with Fact-Checking (*FC*) warning, center row: Machine-Learning (*ML*) warning, and bottom row: Machine-Learning-Accuracy (*MLA*) warning.

its credibility without professional knowledge; 2) the negative effect caused by the fake political news has become a critical issue in our daily life [6]. For example, in the 2016 American presidential election period, a piece of news titled "Pope Francis Shocks World, Endorses Donald Trump for President" [1] shook the world and commoved voters. Therefore, we believe political news should be treated as one of the top priority news types in solving fake news problems.

Procedure. Figure 2 illustrates the flow chart of Experiment 1. Participants were randomly assigned to one of the four conditions. After participants made an informed consent, Phase 1 started. Eight different news (half fake) were presented one at a time in a randomized order. Participants were instructed to view the headline first and then decide whether they have heard about the news (i.e., *Yes, Unsure, No*). Then, participants were asked to judge the accuracy and decide their willingness to share the news on a 5-point Likert scale, respectively (1 means "Very inaccurate" or "I would never share news like this one", 5 means "Very accurate" or "I would love to share news like this one").

After Phase 1, participants completed a demographic questionnaire that asked for age, gender, and etc., as a distraction. Then Phase 2 started, in which participants completed the same three tasks with another set of eight news as Phase 1, except that the warning labels were *removed*. At the end of Phase 2, participants completed additional questions about their computer skill, social media experience, interest in politics, factors that impact their decisions on three tasks, and their trust on the warning on a 5-point Likert scale (1 means they did not trust the warning at all, 5 means they trust the warning a great deal). Phase 3 was conducted one week after Phases 1 and 2. Each participant received emails inviting him/her to evaluate a set of 16 news (half real and half fake) as in Phase 2. The given news included a new set of eight news, and four from Phase 1, and another four from Phase 2. Each participant was compensated for \$0.5 for the completion of Phases 1 and 2, and participants who finished Phase 3 received an extra \$0.5.

[1] https://www.snopes.com/fact-check/pope-francis-donald-trump-endorsement/

3.2 Results From Experiment 1

We recruited 800 MTurk workers on July 27, 2018. After removing nine incomplete submissions, 44 responses with both duplicate GPS coordinates (longitude and latitude provided by Quarltrics) and IP addresses, 178 responses with duplicate GPS coordinates but different IP addresses (rationales adopted from [2]), and 17 responses submitted within 3 minutes (median completion time is about 7 minutes), the numbers of participants that we accepted for the three warning conditions were 132, 136, and 138, respectively. The number of participants recruited in the *CON* condition was 146. In total, 552 participants (55.2% female) were included for data analyses. Participants' average age was 39, with 75% between 20 to 40 years. 55% of participants were college students or professionals who had a bachelor or higher degrees. The demographic distributions were similar among the four conditions.

For our analysis, selection rates of "Yes" for the recognition task were calculated for fake news and real news, respectively. For the detection task, choices of "Very inaccurate" and "Inaccurate" for fake news, and choices of "Accurate" and "Very accurate" for real news, were counted and coded as correct. The selection ratio of "Probably yes" and "I would love to share news like this one" of the sharing task were counted for fake and real news, respectively. For each task, we also measured participants' selection rates of "Unsure" option.

For each phase, specified decision rates (range from 0 to 1) of each participant for each task were transferred into arcsine values, and then entered into 2 (news' legitimacy: *fake*, *real*) × 2 (condition: *CON*, one warning label) mixed analysis of variances (ANOVAs), with a significance level of .05. At Phase 3, we included eight news from Phases 1 and 2, so repetition (*repeated*, *non-repeated*) was added as another within-subject factor for the tests.

Because the proportion of successful fake news *detection* ignores the influence from real news, we also used the SDT examining participants' sensitivity (d') and response bias (c) based on their correct detection of fake news (H) and incorrect detection of real news (FA). To accommodate H and FA rates of 0 or 1, a log-linear correction added 0.5 to the number of H, 0.5 to the number of FA, 1 to the number of signals (fake news), and 1 to the number of noise (real news) [7, 18]. Although the true d' values were underestimated by the log-linear correction [18], the relatively differences across the conditions should reflect differences apparent in the raw accuracy data. Measures of d' and c of detection decisions from Phases 1 and 2 were submitted to two-sample t-tests. At Phase 3, ANOVAs were conducted with repetition added as a within-subject factor.

Phase 1: Effect of warning. Table 1 lists the specified decision rates of each task for each condition in each phase, as well as the SDT measures for the detection task.

Recognition decisions. Across all phases, participants recognized more real news (34.6%) than fake ones (4.6%), $F_s > 99.29, p_s < .001, \eta^2_{ps} > .459$, and were more unsure about the recognition of real news (20.4%) than fake news (7.5%), $F_s > 32.64, p_s < .001, \eta^2_{ps} > .248$. No term involved *condition* was significant except the unsure recognition in *FC* (10.9%) was smaller than in *CON* (14.3%),

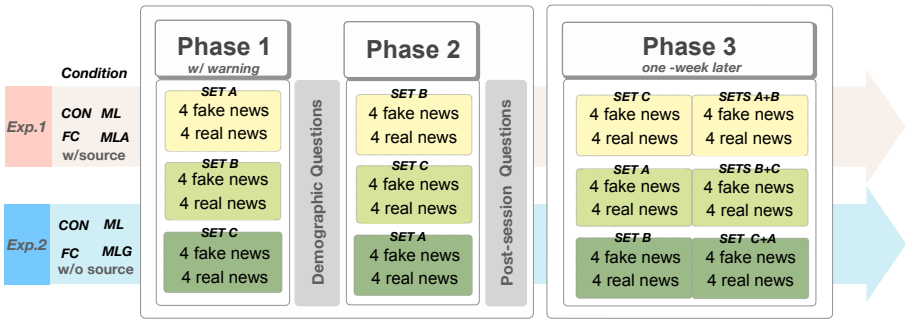

Figure 2: A flow chart showing the experimental design of each phase for both Experiments 1 and 2.

Table 1: Recognition, unsure recognition, correct detection, unsure detection, d', c, sharing, and unsure sharing results of fake and real news of each condition in each phase for Experiments 1 and 2. Sub. means subject, recog. means recognition.

Decision	Exp.1											Exp.2									
	Cond.	Sub. No.	Phase 1		Phase 2		Sub. No.	Phase 3 (New)		Phase 3 (Repeated)		Cond.	Sub. No.	Phase 1		Phase 2		Sub. No.	Phase 3 (New)		Phase 3 (Repeated)	
			Fake	Real	Fake	Real		Fake	Real	Fake	Real			Fake	Real	Fake	Real		Fake	Real	Fake	Real
Recog.	CON	146	4.3%	33.0%	3.9%	31.5%	61	4.9%	31.6%	11.1%	44.7%	CON	153	5.1%	35.8%	5.4%	34.6%	47	2.7%	23.4%	14.9%	39.4%
	FC	132	3.0%	30.3%	3.8%	28.4%	58	2.2%	29.0%	14.7%	45.1%	FC	160	7.5%	33.9%	6.3%	28.4%	60	4.6%	31.3%	16.7%	45.4%
	ML	136	5.9%	38.8%	5.7%	30.3%	58	9.1%	32.3%	17.7%	41.8%	ML	160	4.1%	31.3%	5.2%	29.4%	45	1.1%	25.6%	13.3%	47.2%
	MLA	138	4.0%	28.4%	3.8%	33.2%	50	2.0%	31.0%	9.0%	44.0%	MLG	151	3.6%	37.1%	3.8%	31.5%	54	2.3%	25.0%	13.4%	43.1%
Recog_Unsure	CON	146	8.4%	20.2%	8.7%	24.5%	61	6.6%	21.7%	16.0%	23.4%	CON	153	8.7%	19.6%	9.8%	23.5%	47	9.6%	23.4%	17.0%	27.7%
	FC	132	6.1%	15.7%	9.3%	18.9%	58	5.4%	23.2%	10.3%	19.6%	FC	160	6.1%	21.9%	10.6%	27.8%	60	9.6%	22.9%	15.4%	19.6%
	ML	136	6.4%	19.9%	12.9%	23.5%	58	8.2%	18.1%	18.1%	25.9%	ML	160	8.1%	23.6%	8.9%	25.6%	45	7.2%	20.6%	17.8%	21.7%
	MLA	138	9.1%	25.9%	9.8%	21.7%	50	7.0%	21.0%	15.5%	21.5%	MLG	151	6.8%	19.4%	10.6%	22.5%	54	11.1%	27.8%	19.0%	25.0%
Detection	CON	146	72.6%	39.7%	71.1%	39.7%	61	79.5%	38.9%	60.7%	43.9%	CON	153	70.1%	44.8%	69.0%	41.0%	47	72.3%	38.3%	61.2%	43.6%
	FC	132	79.2%	45.1%	72.0%	40.9%	58	73.7%	42.0%	62.9%	49.1%	FC	160	73.9%	42.5%	68.1%	39.4%	60	70.4%	42.9%	65.0%	51.7%
	ML	136	71.7%	38.4%	65.3%	35.7%	58	67.2%	36.6%	59.1%	42.2%	ML	160	73.8%	39.8%	65.0%	39.5%	45	72.8%	34.4%	63.9%	43.9%
	MLA	138	74.5%	39.7%	72.1%	38.9%	50	72.0%	37.0%	62.5%	47.0%	MLG	151	78.3%	43.0%	67.7%	37.6%	54	76.9%	35.2%	67.1%	43.1%
Detection_Unsure	CON	146	21.6%	35.6%	22.4%	38.7%	61	15.2%	34.8%	27.9%	34.0%	CON	153	21.1%	34.2%	23.4%	40.2%	47	16.5%	39.9%	28.7%	33.0%
	FC	132	13.8%	31.1%	20.3%	34.5%	58	20.5%	36.6%	25.9%	31.3%	FC	160	18.4%	37.2%	22.8%	40.2%	60	21.3%	40.8%	20.4%	29.6%
	ML	136	20.4%	35.8%	27.6%	42.6%	58	23.3%	35.3%	27.2%	35.3%	ML	160	19.4%	37.7%	25.5%	43.3%	45	23.9%	42.8%	28.3%	35.0%
	MLA	138	19.7%	40.4%	22.8%	38.6%	50	21.5%	36.0%	27.0%	33.0%	MLG	151	17.7%	35.8%	27.2%	41.1%	54	18.1%	39.8%	26.4%	30.6%
d'	CON	146	1.17		1.22		61	1.30		0.94		CON	153	1.20		1.22		47	1.20		0.92	
	FC	132	1.37		1.14		58	1.28		1.06		FC	160	1.33		1.19		60	1.34		1.13	
	ML	136	1.12		1.07		58	0.96		0.89		ML	160	1.26		1.17		45	1.21		1.03	
	MLA	138	1.34		1.23		50	1.11		1.05		MLG	151	1.41		1.14		54	1.26		0.97	
c	CON	146	0.04		0.09		61	-0.07		0.22		CON	153	0.11		0.15		47	0.08		0.19	
	FC	132	-0.04		0.04		58	0.06		0.21		FC	160	0.06		0.15		60	0.17		0.21	
	ML	136	0.03		0.16		58	0.06		0.23		ML	160	0.04		0.23		45	0.06		0.18	
	MLA	138	0.07		0.07		50	0.01		0.21		MLG	151	-0.004		0.14		54	-0.03		0.08	
Sharing	CON	146	6.5%	13.7%	6.9%	13.2%	61	9.8%	19.7%	13.1%	20.5%	CON	153	7.8%	15.0%	6.5%	15.0%	47	7.4%	11.2%	7.4%	15.4%
	FC	132	4.5%	15.9%	5.3%	13.6%	58	7.6%	12.1%	9.8%	16.10	FC	160	9.7%	15.2%	10.5%	20.2%	60	11.7%	13.3%	10.8%	15.8%
	ML	136	7.9%	15.4%	7.7%	14.9%	58	11.2%	18.5%	11.2%	13.40	ML	160	5.0%	13.6%	7.0%	12.0%	45	5.6%	8.9%	4.4%	9.4%
	MLA	138	4.7%	14.5%	6.9%	13.2%	50	8.5%	17.0%	5.5%	16.00	MLG	151	5.5%	15.4%	4.5%	12.6%	54	6.5%	12.5%	7.9%	15.3%
Sharing_Unsure	CON	146	6.2%	12.8%	9.1%	14.0%	61	7.4%	16.0%	8.6%	13.9%	CON	153	7.4%	13.1%	8.3%	16.7%	47	5.9%	10.1%	12.2%	11.2%
	FC	132	3.2%	8.0%	6.4%	10.0%	58	4.5%	12.1%	4.9%	15.2%	FC	160	5.6%	14.5%	9.7%	12.2%	60	7.9%	15.4%	6.3%	13.3%
	ML	136	8.8%	16.2%	11.6%	20.2%	58	9.1%	15.1%	10.8%	19.0%	ML	160	7.0%	10.9%	4.8%	12.5%	45	8.3%	9.4%	6.7%	11.1%
	MLA	138	9.2%	16.7%	9.4%	15.6%	50	6.5%	13.5%	13.5%	15.5%	MLG	151	6.8%	12.7%	11.3%	14.1%	54	9.7%	15.3%	7.4%	8.8%

$F_{(1,276)} = 4.31, p = .039, \eta_p^2 = .015$. So we focus on the analyses of detection and sharing decisions in the following parts, but return to recognition decisions in the General Discussion.

Detection decisions. Analyses of correct detection decisions revealed that main effects of news legitimacy were significant across all comparisons, $F_s > 160.56, p_s < .001, \eta_{ps}^2 > .368$. Regardless of conditions, participants correctly detected more fake news (74.4%) than real news (40.7%). Relative to *CON* (56.2%), the overall correct detection rate was higher for *FC* (62%), $F_{(1,276)} = 5.99, p = .015, \eta_p^2 = .021$, but not the other conditions (*ML*: 55%, *MLA*: 57.1%), $F_s < 1.0$. However, the two-way interaction of news legitimacy and the condition was not significant, $F < 1.0$. Thus, the *FC* warning not only increased participants' correct detection of fake news but also increased their correct detection of real news, suggesting that participants may rely on the presence and absence of the warning to judge the legitimacy of news headlines.

Across all comparisons, participants were more unsure in detecting real news (35.8%) than fake news (18.9%), $F_s > 56.92, p_s < .001, \eta_{ps}^2 > .169$, which made sense since the warning label was presented with fake news only. Relative to *CON* (28.6%), only participants in *FC* (22.4%) showed less unsure about their detection, $F_{(1,276)} = 6.05, p = .015, \eta_p^2 = .014$, but not the other conditions (*ML*: 28.1%, *MLA*: 29.9%), $F_s < 1.0$. Also, the reduced unsure detection rate (about 6%) of the *FC* warning was almost equal to the increased correct detection rate of the *FC* warning (about 6%), suggesting that participants relied on the *FC* warning to make decisions mainly when they were uncertain about the news' legitimacy. The main effect of condition did not interact with news legitimacy, $F < 1.0$, indicating the effect of *FC* was similar between fake and real news.

SDT measures. When warning was present, participants showed minimal bias toward detecting news as fake across all conditions (c = 0.02). Compared to *CON* (d' = 1.17), participants' sensitivity to differentiate fake and real news were similar for all warnings (*FC*: d' = 1.37, $t_{(276)} = 1.80, p = .073$; *ML*: d' = 1.12, $t < 1.0$; and *MLA*: d' = 1.34, $t_{(282)} = 1.50, p = .135$).

Sharing decisions. Participants' overall willingness to sharing the news was low (see Table 1), but their willingness to share real news (14.9%) was higher than that of fake news (5.9%), $F_s > 41.78, p_s < .001, \eta^2_{ps} > .130$. Neither the main effect of condition (*CON* vs. *FC* vs. *ML* vs. *MLA*: 10.1% vs. 10.2% vs. 11.7% vs. 9.6%) nor its interaction with news legitimacy were significant, $F_s < 3.51$.

Participants were more unsure about sharing real news (13.5%) than fake news (6.9%), $F_s > 28.89, p_s < .001, \eta^2_{ps} > .095$. Compared to *CON* (9.5%), participants in *FC* (5.6%) were less unsure about their decisions, $F_{(1,276)} = 6.46, p = .012, \eta^2_p = .016$, but not participants in *ML* (12.5%) or *MLA* (12.9%) conditions, $F_s < 1.0$. Consistent with the results of unsure detection decisions, the *FC* warning also reduced participants' uncertainty during sharing decision-making.

Phase 2: Short-term effect of warning. Specified decision rates and SDT measures for Phase 2 are listed in Table 1.

Detection decision. As in Phase 1, the main effect of news legitimacy was significant, $F_s > 149.60, p_s < .001, \eta^2_{ps} > .352$. When the warning was absent in Phase 2, participants' correct detection of fake news (70.1%) was still better than that of real news (38.8%). For unsure option selection, the main effect of news legitimacy was also significant, $F_s > 67.55, p_s < .001, \eta^2_{ps} > .197$. Same as in Phase 1, participants showed less unsure of fake news (23.3%) than that of real news (38.6%). Regardless of the warning's presence or absence, more uncertainty at detecting real news than fake news probably was not due to the lack of decision aid for real news trials. No other terms were significant or approached significance.

SDT measures. When the warning was absent in Phase 2, across all conditions, participants showed similar sensitivity (d' = 1.17) and minimal bias toward detecting news as real (c = 0.09), see Table 1. Neither measures showed difference across conditions, $t_s \leq 1.35$. Taken the results of detection decision and SDT measures together, participants' reasonably accurate detection of fake news but not real news seems mainly due to their uncertainty of real news.

Sharing decision. Without warnings, participants were more willing to share real news (13.7%) than fake news (6.9%), $F_s > 24.25, p_s < .001, \eta^2_{ps} > .079$, and were more unsure about sharing of real news (15%) than fake news (9.1%), $F_s > 12.56, p_s < .001, \eta^2_{ps} > .043$. No term involved condition was significant.

Phase 3: Long-term effect of warning. A total of 225 participants returned for Phase 3. Return rates (*CON*: 41.8%, *FC*: 42.4%, *ML*: 42.7%, *MLA*: 36.2%) and demographics were similar across conditions. Decision results and SDT measures for Phase 3 also are shown in Table 1.

Detection decisions. Correct detection of fake news (67.2%) was still better than that of real news (42.1%), $F_s > 50.46, p_s < .001, \eta^2_{ps} > .305$. And the main effect of news legitimacy interacted with repetition across all comparisons, $F_s > 20.84, p_s < .001, \eta^2_{ps} > .151$. Participants correctly detect more fake news which was presented in Phase 3 only (73.2%) than those from Phases 1 and 2 (61.2%). But an opposite pattern was obtained for the real news: participants correctly detect less real news which was presented in Phase 3 only (38.7%) than those from prior phases (45.4%).

For unsure option selection, both the main effect of news legitimacy and its interaction with repetition were significant across all comparisons, $F_s > 4.03, p_s < .047, \eta^2_{ps} > .033$. Same as prior two phases, participants were more unsure at detecting real news (34.5%) than fake news (23.5%). Besides, participants' uncertainty selection difference between repeated and non-repeated pieces of news was larger for fake news (repeated: 27.0%, non-repeated: 20.0%) than for real news (repeated: 33.4%, non-repeated: 35.7%)

SDT measures. Across conditions, there were no differences for both d' and c for the detection decisions, $F_s < 1.0$. But participants were biased to judge repeated pieces of news as real (c = 0.22) and non-repeated news as fake (c = -0.28), $F_s > 116.58, p_s < .001, \eta^2_{ps} > .517$. Also, participants tended to be less sensitive for repeated news (d' = 0.99) than non-repeated news (d' = 1.16), with the effect of repetition was significant for *FC* and *ML*, $F_s > 3.94, p_s > .049, \eta^2_{ps} > .033$, but not *MLA*, $F_{(1,109)} = 2.95, p = .088, \eta^2_p = .026$.

Sharing decisions. One week later, the willingness to share real news (16.7%) was still higher than that of fake news (9.7%), $F_s > 14.34, p_s < .001, \eta^2_{ps} = .109$. No other effects were significant, except there was a main effect of repetition for the group of *FC*, $F_{(1,115)} = 4.14, p = .044, \eta^2_p = .035$. Participants' willingness to share news was reduced for *FC* (11.4%) than for *CON* (15.8%). For unsure option, only the main effects of news legitimacy were significant, $F_s > 16.47, p_s < .001, \eta^2_{ps} = .131$. Again, participants showed more unsure to share real news (15%) than fake news (8.1%).

Post-session questions. 72.6% participants did not have a major or work experience in computer-related fields, and 97.5% of participants did not show concern about using computers successfully in diverse situations. 73.2% of participants indicated that they used social media, such as Facebook and Twitter, daily or a few times a week. 82.6% of participants had an interest in politics.

When asked participants to confirm factors that impact their decisions on news' credibility and sharing on social media, Most participants selected source as the most influential factor for their detection (59.2%) and sharing (46.7%) decisions. Overall, participants did not show much trust on warnings, with 31.8% gave "a great deal" or "a lot" trust, 30.8% indicated their trust was moderate, and 37.4% showed a little or no trust. A chi-squared test showed that participants' trust on warning varied across conditions, $\chi^2_{(2)} = 7.27$, $p = .026$, mainly due to more trust obtained for *FC* (40.2%) than *ML* (25%), $p_{adj.} = .023$.

3.3 Discussion

In Experiment 1, we proposed two machine-learning warning and evaluated their effects and one fact-checking warning in help individuals mitigate fake news. In Phase 1, relative to *CON* in which no warning was present, better detection results were obtained for the *FC* warning but not the *ML* and *MLA* warnings. The *FC* warning improved the correct detection of both fake and real news, suggesting that participants may use the presence and absence of warning as the criterion to make their detection decision, which is in agreement with the more trust obtained for the *FC* warning in post-session questions. When no warnings were displayed with fake news in Phases 2 and 3, the effect of *FC* disappeared. The *FC* warning did not show any short-term or long-term effect in helping participants detect fake news, probably because there were no details to inform participants about why the fake news was labeled. Although machine learning is a buzzword, participants showed less trust on the two machine learning warnings than the fact-checking

warning, suggesting that they may not necessarily understand what it is and consequently, distrust its use for fake news detection.

4 EXPERIMENT 2

In Experiment 2, to increase the transparency of machine learning algorithms, we proposed a Machine-Learning-Graph (*MLG*) warning in which factors that a machine learning algorithm considers during the fact checking are provided under "Disputed by a Machine Learning Algorithm" label. Because participants identified the news source as the most influential factor in their judgment of the news headlines' legitimacy, we also assessed the robustness of the effect of the *FC* warning from Experiment 1 by removing the source information. We also included *CON* and *ML* without sources to provide baselines for evaluation.

4.1 Participants, Materials,Procedure

We recruited extra 800 MTurk workers on October 16, 2018. The requirements to participate in this study was same as Experiment 1. Furthermore, any participants who already participated in the previous study were excluded.

Materials and procedures of Experiment 2 were identical to Experiment 1 except as noted. First, we removed the source for all the 24 news headlines used in Experiment 1. Second, for the *MLG* condition, we added an extra bar chart below the warning label to represent factors that our hypothetical multi-modal machine learning algorithm considers (e.g., [20, 25, 33, 40]). Three factors, "Source Reliability", "Content Trustfulness", and "Picture/Video Truthfulness", were listed from top to bottom. A filled bar graph was accompanying each factor, and the length of each bar indicates values that the machine learning algorithm derived for the evaluation of the factor. The shorter the filled blue bar, the less reliable or accuracy for the news (see Figure 3).

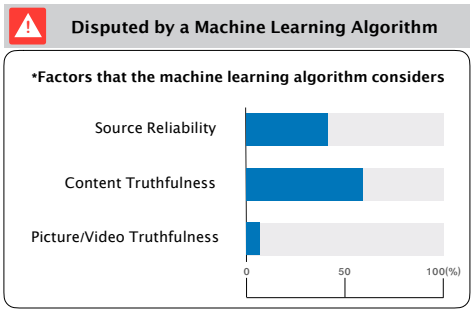

Figure 3: Machine-Learning-Graph (*MLG*) warning of Experiment 2.

4.2 Results From Experiment 2

Using the same criterion as Experiment 1, we got a total of 624 (54.9% female) valid responses, with 153, 160, 160, and 151, for *CON*, *FC*, *ML*, and *MLG*, accordingly. Participants' average age was 39.5 years. 54.2% of participants hold a bachelor or higher degree. For each task in each phase, specified decision rates and SDT measures as a function of the condition were calculated for each participant. Analyses of the decisions rates and SDT measures were conducted in the same way as Experiment 1.

Phase 1: Effect of warning. Table 1 lists the specified decision rates and SDT measures. Same as Experiment 1, participants recognized more real news (34.5%) than fake news (5.1%) regardless of conditions or phases, $F_s > 103.94, ps < .001, \eta^2_{ps} > .512$, and were more unsure about recognizing real news (21.1%) than fake news (7.4%), $F_s > 19.83, ps < .001, \eta^2_{ps} > .181$. Again, we focus on the analyses of detecting and sharing decisions in the following parts, but return to recognition decisions in the General Discussion.

Detection decisions. The main effect of news legitimacy was significant across all comparisons, $F_s > 130.56, ps < .001, \eta^2_{ps} > .296$. Participants correctly detected more fake news (74.0%) than real news (42.5%). Moreover, for *MLG*, compared to *CON*, there was a two-way interaction of news legitimacy and condition, $F_{(1,302)} = 5.48, p = .020, \eta^2_p = .018$. Those participants made more correct decisions on fake news with the *MLG* warning (78.3%) than without warning (70.1%), but their correct decision on real news was similar between the two conditions (*CON*: 44.8%, *MLG*: 43.0%), suggesting the effectiveness of *MLG* in reducing participants' fake news susceptibility. Relative to *CON*, neither the main effect of condition nor its interaction with the condition was significant for the correct detection with *FC* or *ML* warnings, $F_s < 3.05$.

For unsure detection, compared to the *CON*, only the main effect of news legitimacy was significant across all comparisons, $F_s > 74.86, ps < .001, \eta^2_{ps} > .194$, Participants were more uncertain about the accuracy of real news (36.2%) than fake news (19.2%).

SDT measures. When warning was present, compared to *CON* (d' = 1.20), participants' sensitivity to differentiate fake and real news was better for *MLG* (d' = 1.41), $t_{(302)} = 1.98, p = .048$, but not other conditions (*FC*: d' = 1.33, $t_{(311)} = 1.17, p = .242$, and *ML*: d' = 1.26, $t < 1.0$). Relative to *CON* (c = 0.11), participants showed similar bias for each warning [*MLG* (c = -0.004): $t_{(302)} = -1.84, p = .067$; *FC* ($c$ = 0.06): $t < 1.0$; and and *ML* (c = 0.04): $t_{(311)} = -1.05, p = .294$].

Sharing decisions. Compared to *CON*, only the main effect of news legitimacy was significant for both sharing and unsure decisions for all warnings, $F_s > 15.31, ps < .001, \eta^2_{ps} > .047$. In general, participants were more willing to share real news (14.8%) than fake news (7.0%), and they also showed more uncertainty at sharing real news (12.8%) than fake news (6.7%).

Phase 2: Short-term effect of warning. Decision results and SDT measures of each task are shown in Table 1.

Detection decisions. When the warning was absent after a short distraction task, the main effect of news legitimacy was still significant across all comparisons, $F_s > 119.49, ps < .001, \eta^2_{ps} > .278$. Participants correct detection of fake news (67.4%) was better than that of real news (39.4%). However, the effect of *MLG* obtained in Phase 1 disappeared, $F < 1.0$. For unsure option selection, participants still showed more unsure for real news (41.2%) than fake news (24.7%) across all comparisons, $F_s > 68.11, ps < .001, \eta^2_{ps} > .184$.

SDT measures. When the warning was absent, neither measures showed difference across conditions, $t_s \leq -1.29, ps \geq .197$.

Sharing decisions. Sharing decisions also showed the same pattern as prior results (see Table 1). Participants were more willing to share real news (15.0%) than fake news (7.2%), $F_s > 36.69, ps < .001, \eta^2_{ps} > .106$. For unsure option selection, participants also showed more uncertainty about sharing real news (13.8%) than fake news (8.5%), $F_s > 18.69, ps < .001, \eta^2_{ps} > .058$. Moreover, the

effect of warning was revealed in all comparisons. Relative to *CON*, participants who saw the *MLG* warning in Phase 1, increased their uncertainty about sharing fake news but reduced their uncertainty about sharing real news, $F_{(1,302)} = 4.14, p = .043, \eta_p^2 = .014$. Participants who saw the *FC* warning in Phase 1 showed the similar pattern as participants in *MLG*, $F_{(1,311)} = 3.87, p = .050, \eta_p^2 = .012$. But their increased susceptibility of fake news was numerically smaller than that of *MLG* and reduced susceptibility of real news was numerically larger than that of *MLG*. And for participants in *ML*, they reduced their uncertainty of sharing both fake and real news, $F_{(1,311)} = 4.61, p = .033, \eta_p^2 = .015$.

Phase 3: Long-term effect of warning. After one week, 206 participants returned for Phase 3. Return rates (*CON*: 30.7% , *FC*: 37.5%, *ML*: 28.1%, *MLG*: 35.8%) and demographics were similar across conditions. Decision results were also shown in Table 1.

Detection decisions. Same as Experiment 1, participants still correctly detected more fake news (68.8%) than real news (41.9%) one week later, $F_s > 37.17, ps < .001, \eta_{ps}^2 > .261$, and their correct detection pattern varying as a function of repetition, $F_s > 9.76, ps < .002, \eta_{ps}^2 > .098$. Across all conditions, participants' correct detection of repeated fake news (64.4%) was smaller than their correct detection of non-repeated fake news (73.1%). However, participants correctly detected more repeated real news (45.9%) than non-repeated real news (38.0%). Although participants in the *MLG* condition showed numerically better results in detecting fake news, the long-term effects of *MLG* were not significant, $F_s < 1.0$.

Participants were more unsure about the selection of real news (36.4%) than fake news (23.0%), $F_s > 18.80, ps < .001, \eta_{ps}^2 > .173$. Although the main effect of repetition was not significant, it interacted with the news legitimacy, $F_s > 10.41, ps < .002, \eta_{ps}^2 > .104$. Participants were more unsure about detecting fake news from real news which was repeated than those which were non-repeated.

SDT measures. Same as Experiment 1, there were no differences for both d' and c for the detection decisions across conditions, $F_s < 1.0$. Nevertheless, participants showed less sensitivity for the repeated news headlines ($d' = 1.01$) than for the non-repeated news headlines ($d' = 1.26$), $F_s > 4.61, ps < .035, \eta_{ps}^2 > .045$. They also tended to be biased to judge repeated news as real ($c = 0.16$) than non-repeated news ($c = 0.07$), with the effect of repetition was significant for *ML* and *MLG*, $F_s > 5.55, ps < .021, \eta_{ps}^2 > .058$, but not *FC*, $F_{(1,105)} = 3.51, p = .064, \eta_p^2 = .032$.

Sharing decisions. Same as prior phases, participants showed more willingness to share real news (12.6%) than fake news (8.0%), $F_s > 8.42, ps < .005, \eta_{ps}^2 > .074$. Participants only showed more unsure about sharing real news than fake news for the comparison between *CON* and *FC*, $F_{(1,105)} = 9.95, p = .002, \eta_p^2 = .087$. The two-way interaction of repetition and condition was significant for the comparison between *CON* and *MLG*, $F_{(1,99)} = 8.62, p = .004, \eta_p^2 = .080$. Compared to *CON*, participants in *MLG* condition showed more uncertainty at sharing news that was non-repeated but less uncertainty at sharing news that were repeated.

Post-session questions. Overall results of post-session questions in Experiment 2 were similar to those from Experiment 1. 72.4% of participants did not have a major or work experience in computer-related fields, and 98.2% of participants did not concern about using computers successfully in diverse situations. 74.2% of participants indicated that they used social media, such as Facebook and Twitter more than a few times a week, and 84.8% of participants had an interest in politics. When asked how much their trust on the warning when evaluating the accuracy of news during the study, participants did not show much trust on warnings in general, with 16.8% gave "a great deal" or "a lot" trust, 28.2% indicated their trust was moderate, and 55% showed a little or no trust. Participants' trust level also varied across warnings, $\chi_{(2)}^2 = 34.40, p < .001$. Participants showed more trust for *FC* (30.6%) than *ML* (7.5%), $p_{adj.} < .001$, and *MLG* (11.9%), $p_{adj.} < .001$, respectively.

4.3 Discussion

After removing source within each news headlines at Experiment 2, we did not obtain the effect of the *FC* warning as in Experiment 1. Compared to *CON* in which no warning was presented, the *MLG* warning improved participants' detection of fake news and increased their sensitivity to differentiate fake and real news while the *ML* warning did not. When warnings were absent in Phases 2 and 3, neither the main effect of warning nor its interaction with other factors were significant for detection decisions. However, the effect of *MLG* and *FC* were revealed in participants' increased uncertainty of sharing fake news but reduced uncertainty in sharing real news in Phase 2, suggesting a short-term effect for both warnings. Although participants showed better fake news detection with *MLG* in Phase 1, their trust on the *MLG* warning was less than that of the *FC* warning, suggesting that participants' better detection of fake news with *MLG* in Phase 1 was mostly due to their reliance on the factors that presented within the warning.

5 GENERAL DISCUSSION

Across two experiments, we proposed three machine-learning warnings and evaluated their effects and a fact-checking warning in helping individuals mitigate fake news. Both decision rates and SDT measures showed the effect of *MLG* warning in helping participants differentiate fake news from real ones. When no warnings were displayed in Phase 2, although the *MLG* warning did not impact individuals' detection decisions, participants increased their uncertainty in sharing fake news but reduced their uncertainty in sharing real news, suggesting a short-term effect of the warning.

We obtained that the effect of *FC* warning increased participants' correct detection of both fake and real news when the source was included in news headlines but not when sources were excluded. Although the *FC* warning did not impact individuals' detection decisions when the source was excluded, they increased participants' uncertainty in sharing fake news and reduced their uncertainty in sharing real news when the warning was not displayed in Phase 2, suggesting a short-term effect. Thus, our results did not replicate [9], but are somewhat consistent with [27], showing a small effect of the warning. With the *FC* warning, participants not only increased the correct detection of fake news to which the warning was attached but also the correct detection of real news, suggesting that participants probably relied on the presence and absence of the warning to make the detection decision.

5.1 Limited Effect of Warning Labels

All the warnings that have been implemented in current and prior studies (e.g., [27]) revealed a small effect on mitigating the fake

news. One possible reason is that all those proposed warnings are passive, which indicate misinformation to participants without interrupting their primary task. i.e., viewing news headlines and obtaining new information. Prior studies on cybersecurity, e.g., phishing [19, 22], showed that participants ignored passive security indicators and relied instead mainly on the website contents to decide the trustworthiness of a web page. The results of current Experiment 1 showed a similar pattern, in that participants mainly relied on the source of news to make the news' legitimacy decisions even when the warning labels were present. Therefore, one way to improve the effectiveness of warning is to make it active, which will capture users' attention and force users to choose one of the options that were presented by the warning [12, 14, 41].

However, a zero-day exploit of fake news will leave no opportunity for automatic detection and prevention, and people need to make a decision on their own [29]. Therefore, the ability to tell fake news from real ones is an important skill for individuals to acquire. Training is one promising approach to address individuals' inability to differentiate fake and real news. Also, prior studies in cybersecurity provided evidences that knowledge gained from training enhanced the effectiveness of phishing warnings [43]. Therefore, another way to improve the effect of warning is to embed training within the warning and use each warning as an opportunity to train users on how to mitigate fake news.

5.2 Better Recognition of Real News

A point to note about the present study is that overall participants recognized more real news than fake news, and also showed more uncertainty at recognizing real news than fake news. "Recognition" and "unsure" decisions represent two distinct processes for recognition memory, *recollection*, and *familiarity* [21]. The distinction is that people could recognize a piece of news as familiar but not being able to recollect where he or she previously saw it.

Across three phases, participants appeared reasonably accurate at detecting fake news, but their correct detection of real news was less than chance. SDT measures did not show that participants were biased in judging news as fake, thus the poor detection of real news was mainly due to participants' more uncertainty about detecting real news than fake news. A further Pearson correlation analysis revealed that the unsure recognition of real news had a statistically significant positive linear relationship with the unsure detection of real news for both experiments, $p_s < .001$. The strength of the association was approximately moderate for Experiment 1, $r = .294$, and there was a small correlation $r = .245$ for Experiment 2.

Consistent with [17], our results showed that participants' willingness to share news was low in general and was lower for fake news than real news. Moreover, our study revealed that participants were more uncertain about sharing real news than fake news. For both experiments, Pearson correlation analysis showed a significant positive association between unsure recognition and the unsure sharing, $p_s < .003$, but with a small correlation, $r = .268$ for Experiment 1, and $r = .119$ for Experiment 2.

Altogether, the better recognition and more uncertainty of real news suggest that participants may have been exposed to those pieces of real news previously, and their familiarity (uncertainty) with real news seems to have impaired their evaluation of news' accuracy and their sharing decisions.

5.3 Effect of Repetition

At Phase 3, for those pieces of news that were repeated, participants showed better recognition. Moreover, the increase of recognition was more evident for real news than fake news, suggesting the repetition increased more recollection of real news than fake news. Consistent with the effect of repetition obtained by [27], SDT measures further revealed that participants were less sensitive and more biased to judge news headlines as real for news from prior phases than the news that presented in Phase 3 only. Participants' unsure detection was also increased for the repeated pieces of news, however, the increase was more evident for fake news than real news, suggesting that the repetition mainly increased participants' familiarity (uncertainty) of fake news. Therefore, our study provided evidence that the repetition probably impacts the detection of fake and real news differently.

Human memory has been described an optimization of information retrieval, which uses the statistics derived from past experience to estimate which knowledge will be currently relevant [1]. Besides allowing individuals remembering objects and events that they have actual experience, human memory systems are subject to distortion, bias, and the creation of illusions [23, 32]. Combining the overall better recognition of real news, increased recollection of repeated real news and increased familiarity with repeated fake news, our study further indicates the important role that memory plays in individuals' belief in fake news. Further research should be conducted to explore the extent to which memory affects individuals' belief in fake news.

5.4 Limitations

In our experiments, the effectiveness of warning labels was evaluated with a convenience sample of Amazon MTurk workers, who tended to be young, more educated, and more tech-savvy than the general public. Thus, the generalization of current findings to participants with different demographics needs to be further examined. In addition, the experiment design is limited in its ecological validity. We considered a more ecologically valid method, such as providing social media interface during the study, but we decided to present news headlines to exclude extraneous variables that may have an effect on the outcomes, which increased our confidence of the internal validity of obtained results. Note that such a design was the same as the prior studies [9, 27], which made our results comparable to the prior ones.

Another possible confound was that participants may have experienced the fact-checking warning previously but not the machine learning warnings. Better performance only obtained for the *MLG* warning but not *ML* and *MLA* warnings indicate that novelty may be not critical. Finally, all news headlines in our study are politically related, so generalizing the findings to other types of misinformation needs to be further investigated. Finally, in this study, we did not consider participants' political stance as a factor due to our main interest in warning labels and prior study showed that the partisan bias did not significantly affect participants' susceptibility to fake

news [28]. However, recently Gao et al. [15] obtained results indicating that the warning label are more effective for participants in the liberal group than participants in the conservative group. Therefore, to understand whether pre-existing political stance interacts with a machine-learning warning, the future studies can consider political stance of participants as an extra factor and measure how it impacts participants' belief in fake news.

6 CONCLUSION

In this work, we conducted two online experiments to understand the impact of machine-learning warning on reducing individuals' fake news susceptibility. Each experiment consisted of three phases examining participants' recognition, detection, and sharing of fake news, respectively. Across three machine-learning warnings, the Machine-Learning-Graph warning increased participants' sensitivity in differentiating fake from real news, but participants showed limited trusted on it. Our study results imply that a transparent machine learning algorithm (that explains the detail results) may be critical to improving individuals' fake news detection but not necessarily to increase their trust.

7 ACKNOWLEDGEMENT

This work was in part supported by NSF awards #1742702, #1820609, and #1915801, and ORAU-directed R&D program in 2018.

REFERENCES

[1] John R Anderson and Robert Milson. 1989. Human memory: An adaptive perspective. *Psychological Review* 96, 4 (1989), 703–719.
[2] Hui Bai. 2018. Evidence that a large amount of low quality responses on MTurk can be detected with repeated GPS coordinates. (2018). https://goo.gl/19KCHG.
[3] David Bawden and Lyn Robinson. 2009. The dark side of information: Overload, anxiety and other paradoxes and pathologies. *J. of Information Science* 35, 2 (2009), 180–191.
[4] Gaurav Bhatt, Aman Sharma, Shivam Sharma, Ankush Nagpal, Balasubramanian Raman, and Ankush Mittal. 2018. Combining neural, statistical and external features for fake news stance identification. In *The Web Conf. (WWW)*. 1353–1357.
[5] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 1–12.
[6] Carole Cadwalladr. 2017. The great British Brexit robbery: how our democracy was hijacked. https://tinyurl.com/lkhgkdk. (2017). Accessed: 2019-01-10.
[7] Casey Inez Canfield, Baruch Fischhoff, and Alex Davis. 2016. Quantifying phishing susceptibility for detection and behavior decisions. *Human Factors* 58, 8 (2016), 1158–1172.
[8] Michaela Cavanagh. 2018. Climate change: 'Fake news', real fallout. (2018). https://goo.gl/tCbwYq Accessed: 2019-01-10.
[9] Katherine Clayton, Spencer Blair, Jonathan A Busam, and et al. 2019. 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* (2019), 1–23. https://doi.org/10.1007/s11109-019-09533-0
[10] Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. In *78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, Vol. 52. 1–4.
[11] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings. In *Privacy Enhancing Technologies*. 92–112.
[12] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. 2008. You've been warned: An empirical study of the effectiveness of web browser phishing warnings. In *ACM CHI*. ACM, 1065–1074.
[13] Craig Silverman et al. 2016. Hyperpartisan Facebook pages are publishing false and misleading information at an alarming rate. (2016). https://goo.gl/6pWtTT
[14] Adrienne Porter Felt, Alex Ainslie, Robert W Reeder, Sunny Consolvo, Somas Thyagaraja, Alan Bettes, Helen Harris, and Jeff Grimes. 2015. Improving SSL warnings: Comprehension and adherence. In *ACM CHI*. ACM, 2893–2902.
[15] Mingkun Gao, Ziang Xiao, Karrie Karahalios, and Wai-Tat Fu. 2018. To label or not to label: The effect of stance and credibility labels on readers' selection and perception of news articles. *ACM CHI* 2, CSCW (2018), 55.
[16] David M Green and John A Swets. 1966. *Signal detection theory and psychophysics*. Wiley, New York, NY.
[17] Andrew Guess, Jonathan Nagler, and Joshua Tucker. 2019. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances* 5, 1 (2019), eaau4586.
[18] Michael J Hautus. 1995. Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior Research Methods, Instruments, & Computers* 27, 1 (1995), 46–51.
[19] Amir Herzberg and Ahmad Gbara. 2004. *Trustbar: Protecting (even naive) web users from spoofing and phishing attacks*. Technical Report. Cryptology ePrint Archive, Report 2004/155. http://eprint. iacr. org/2004/155.
[20] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *ACM Multimedia Conf.* 795–816.
[21] Colleen M Kelley and Larry L Jacoby. 2000. Recollection and familiarity: Process-dissociation. In *The Oxford handbook of memory*, Endel E. Tulving and Fergus I. M. Craik (Eds.). Oxford University Press, New York, 215–228.
[22] Eric Lin, Saul Greenberg, Eileah Trotter, David Ma, and John Aycock. 2011. Does domain highlighting help people identify phishing sites?. In *ACM CHI*. ACM, 2075–2084.
[23] Elizabeth F Loftus. 2005. Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory* 12, 4 (2005), 361–366.
[24] Neil A Macmillan and Douglas C Creelman. 2004. *Detection theory: A user's guide*. Lawrence Erlbaum, Mahwah, NJ.
[25] Shivam B Parikh and Pradeep K Atrey. 2018. Media-rich fake news detection: A survey. In *IEEE Conf. on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 436–441.
[26] Frank Pasquale. 2015. *The black box society: The secret algorithms that control money and information*. Harvard University Press, Cambridge, MA.
[27] Gordon Pennycook, Tyrone Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *J. of Experimental Psychology: General* 147, 12 (2018), 1865–1880.
[28] Gordon Pennycook and David G Rand. 2018. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* (2018). https://doi.org/10.1016/j.cognition.2018.06.011
[29] Robert W Proctor and Jing Chen. 2015. The role of human factors/ergonomics in the science of security: decision making and action selection in cyberspace. *Human Factors* 57, 5 (2015), 721–727.
[30] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. 2931–2937.
[31] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *ACM SIGKDD int'l conf. on knowledge discovery and data mining (KDD)*. ACM, 1135–1144.
[32] Henry L Roediger III and Kathleen B McDermott. 2000. Tricks of memory. *Current Directions in Psychological Science* 9, 4 (2000), 123–127.
[33] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *ACM Conf. on Information and Knowledge Management (CIKM)*. ACM, 797–806.
[34] Scott Shane. 2017. From headline to photograph, a fake news masterpiece. (2017). https://goo.gl/tmiw7s
[35] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.
[36] Kai Shu, Suhang Wang, and Huan Liu. 2018. Understanding user profiles on social media for fake news detection. In *IEEE Conf. on Multimedia Information Processing and Retrieval (MIPR)*. 430–435.
[37] John A Swets. 1964. *Signal detection and recognition in human observers: Contemporary readings*. Wiley, New York, NY.
[38] Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506* (2017).
[39] Sander Van der Linden, Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach. 2017. Inoculating the public against misinformation about climate change. *Global Challenges* 1, 2 (2017).
[40] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 849–857.
[41] Min Wu, Robert C Miller, and Simson L Garfinkel. 2006. Do security toolbars actually prevent phishing attacks?. In *ACM CHI*. ACM, 601–610.
[42] Aiping Xiong, Robert W Proctor, Weining Yang, and Ninghui Li. 2017. Is domain highlighting actually helpful in identifying phishing web pages? *Human Factors* 59, 4 (2017), 640–660.
[43] Aiping Xiong, Robert W Proctor, Weining Yang, and Ninghui Li. 2018. Embedding training within warnings improves skills of identifying phishing webpages. *Human Factors* (2018). https://doi.org/10.1177/0018720818810942