An Experimental Study to Understand User Experience and Perception Bias Occurred by Fact-checking Messages

Sungkyu Park
Institute for Basic Science (IBS)
Daejeon, South Korea
shaun01.park@gmail.com

Jaimie Yejean Park
Samsung Electronics
Suwon, South Korea
jaimieee@gmail.com

Hyojin Chin
Institute for Basic Science (IBS)
Daejeon, South Korea
tesschin@gmail.com

Jeong-han Kang
Department of Sociology,
Yonsei University
Seoul, South Korea
jhk55@yonsei.ac.kr

Meeyoung Cha
Korea Advanced Institute of Science
and Technology (KAIST) & IBS
Daejeon, South Korea
meeyoungcha@kaist.ac.kr

ABSTRACT

Fact-checking has become the de facto solution for fighting fake news online. This research brings attention to the unexpected and diminished effect of fact-checking due to cognitive biases. We experimented (66,870 decisions) comparing the change in users' stance toward unproven claims before and after being presented with a hypothetical fact-checked condition. We found that, first, the claims tagged with the 'Lack of Evidence' label are recognized similarly as false information unlike other borderline labels, indicating the presence of *uncertainty-aversion bias* in response to insufficient information. Second, users who initially show disapproval toward a claim are less likely to correct their views later than those who initially approve of the same claim when opposite fact-checking labels are shown — an indication of *disapproval bias*. Finally, user interviews revealed that users are more likely to share claims with Divided Evidence than those with Lack of Evidence among borderline messages, reaffirming the presence of uncertainty-aversion bias. On average, we confirm that fact-checking helps users correct their views and reduces the circulation of falsehoods by leading them to abandon extreme views. Simultaneously, the presence of two biases reveals that fact-checking does not always elicit the desired user experience and that the outcome varies by the design of fact-checking messages and people's initial view. These new observations have direct implications for multiple stakeholders, including platforms, policy-makers, and online users.

CCS CONCEPTS

• **Applied computing → Sociology**; • **Human-centered computing → Interaction design theory, concepts and paradigms**; *Interaction design process and methods.*

KEYWORDS

Social media, fact-checking, perception, risk-aversion, uncertainty-aversion, disapproval bias

ACM Reference Format:
Sungkyu Park, Jaimie Yejean Park, Hyojin Chin, Jeong-han Kang, and Meeyoung Cha. 2021. An Experimental Study to Understand User Experience and Perception Bias Occurred by Fact-checking Messages. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3442381.3450121

1 INTRODUCTION

Fake news and online misinformation have become a considerable challenge in this digital era. As more people rely on social network services and online media to search information and discover news, the risk of their exposure to misinformation through these platforms has increased [13, 43]. False and incorrect information has shown to skew asset prices [4], lead people in a disaster situation to panic [20], threaten democratic outcomes [45], and grow cynicism towards politics and journalism [2].

Fact-checking is a significant and feasible solution to fight against online misinformation [7]. These systems help people recognize fake news and pay more attention to questionable features on news headlines [6]. The desired effects of online fact-checking systems are two-fold. First is to stop the diffusion of misinformation and second is to help users have the correct viewpoint and an accurate understanding of a claim [1].

However, what if fact-checking systems cannot label a claim as either true or false? This is because claims in the real world do not always lend themselves to simple true-or-false verdicts [33]. Fact-check services already offer indecisive, borderline labels (e.g., 'Mixture' and 'Unproven' for Snopes, 'half-true' for PolitiFact), yet these messages may cause unintentional side effects like drawing back to their existing view [28]. One study has demonstrated that exposure to a general warning lowered the perceived accuracy of true information online [9]. Not knowing in advance, such unintended consequences may significantly affect reducing a future fact-checking system's credibility.

This research extends our previous work, titled "The presence of unexpected biases in online fact-checking," published in [33]. Here, we had investigated how the effect of fact-checking could

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3450121

change by various factors and had shown that fact-checking helps self-correct one's views, yet its effect is weaker for individuals who perceived the claim negatively at first. The study also highlighted the unexpected role of wording; borderline fact-check messages showed different influences on developing a user stance toward online claims.

The current study offers additional quantitative findings from the same randomized survey, explaining the two relevant cognitive biases (i.e., uncertainty-aversion and disapproval bias) in depth. We newly show how the stance difference between one's initial views on claims (hereafter 'pre-stance') and views after seeing fact-checking messages (i.e., the succeeding stance after being exposed to the fact-checking conditions, hereafter 'post-stance') affects the viewer's willingness to share the message. As a way to quantify the efficacy of fact-checking messages, we explore whether users react differently with the combination of fact-checking labels and their pre-stance towards given claims. This experiment involved a set of randomized surveys with 11,145 individuals.

We also provide a qualitative explanation for some of these novel findings by conducting semi-supervised interviews to learn the user experience and perception of different fact-checking conditions. The interviews reaffirm the positive effects of fact-checking in leading people to rethink their initial view, i.e., people become less prone to maintaining radical attitudes and instead lean toward a neutral position upon seeing a fact-checking label. The same conclusion could be drawn from our quantitative study, which showed fact-checking corrects people's opinions on rumors, i.e., 38.5% of people self-corrected their views upon seeing an opposite fact-checking label to their pre-stance. The user interviews showed people's perspectives sharing unproven claims; similar to quantitative findings, participants mentioned they are more likely to share claims that are marked with the Divided Evidence label (than those with Lack of Evidence). This reaffirms the uncertainty-aversion bias. While not brought out explicitly from the interviews, our quantitative analysis had shown disapproval bias for individuals with negative pre-stance; self-correction is less likely to happen when pre-stance is negative. These findings help us understand how online users experience and perceive fact-checking labels in the wild.

We make the survey design and the gathered responses publicly available for the research community. The complete questionnaire and the refined dataset with descriptions are made available.[1]

2 BACKGROUNDS

2.1 Effort to Reduce Misinformation in Online

Several studies from computer science have presented algorithms to automatically detect misinformation [10, 39], whereas studies from social sciences have investigated various human factors that affect people's beliefs in information propagation [42, 43]. If the former contributes to engineering efforts at the platform level, the latter contributes to decision-making efforts at the policy level, and these two categories (i.e., structural changes and empowering individuals) are the significant interventions currently studied [27]. In particular, several studies proposed to build a classifier that determines whether the given news article is a hoax based on data characteristics. This classification will identify a smaller set of news articles (out of millions produced every day) and determine which ones are likely false and therefore need to be fact-checked.

Studies have identified the temporal, topological, and semantic characteristics of misinformation from millions of tweets and have built a classifier to detect rumors [26, 30]. Another study characterized the traits of long spread hoax articles in Wikipedia [25]. Likewise, much effort has been on developing software that detects false information automatically [29, 35, 38] and on devising models utilizing deep embeddings like BERT to extract structured information or to rank claim check-worthiness [23, 47]. Some studies focused on modeling probabilistic patterns describing users flagging behaviors (i.e., crowdsourcing) on suspicious news [24].

What we are interested in, for this study, is whether such systems effectively correct users' perceptions and calming the spread [34]. For instance, Facebook's design change for abandoning the "Disputed" label on user-self-reported news posts demonstrates that fact-checking labels could sometimes amplify the circulation of false stories [40]. One study ran an online experiment with 132 Internet users and subsequently found that showing stance labels was more effective in fighting falsehoods than showing credibility labels when it comes to political news stories [17]. Another research considered fact-checking as a kind of social activity and found that fact-checking interventions are mainly delivered by strangers, whereas they got responses if friends provoked [22].

2.2 Cognitive Bias for Borderline Fact-Check Condition

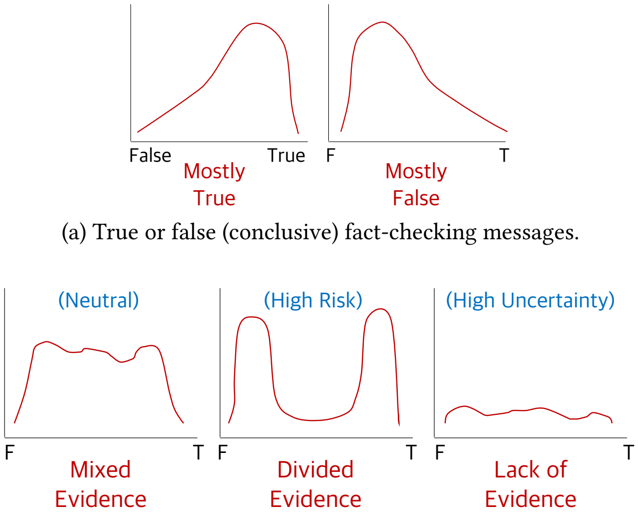
In this paper, we try to differentiate between two similar psychological processes — risk aversion and uncertainty aversion — in estimating the borderline fact-checking condition's effect. *Risk* refers to situations where probabilities can formulate the perceived likelihoods of events of interest [12]. For example, large-scale adventurous research implicates that gigantic success and failure are both highly expected. In a similar sense, a borderline fact-check decision is regarded as "high risk" if both true and false verdicts are highly probable simultaneously. When fact-checkers' decisions on a claim are clearly divided between Definitely True and Definitely False, betting on either true or false also implies a high risk of failure, i.e., Divided Evidence in Figure 1.

On the contrary, *uncertainty* refers to situations where the information available to the person who makes decisions is too vague to be encapsulated by a probability measure [12]. The risk is not evident under uncertainty, and, thus, the expected degree of failure or success is unclear. Suppose the system fails to collect reliable data (Lack of Evidence in Figure 1), such as when fact-checkers are unfamiliar with a given claim and are unwilling to conclude it. In that case, a fact-checked decision is linked with "high uncertainty."

Finally, it is considered truly *neutral*, if fact-checking decisions are not associated with both high risk and high uncertainty, such as fact-checkers present a full spectrum of choices from Definitely True to Definitely False on a single claim (i.e., Mixed Evidence) [33].

The concept of our alternative forms of borderline fact-check condition, i.e., Mixed Evidence, Divided Evidence, and Lack of Evidence, was derived from the rating standards of fact-checking outlets such as Snopes, PolitiFact, AllSides, and Washington Post Fact

[1] Data released via GitHub. https://github.com/dscig/rumorstance.



(a) True or false (conclusive) fact-checking messages.

(b) Borderline fact-checking messages.

Figure 1: Designs of the five hypothetical fact-checking conditions from [33]. The bottom three plots imply borderline verdicts: 1) Mixed Evidence represents a case when fact-checkers gave the full spectrum of true to false decisions, 2) Divided Evidence represents when fact-checkers sharply divided in their decisions, and 3) Lack of Evidence represents when there is not enough information for judgment. The blue text description is the latent representation linked to each condition and was not shown to the users.

Checker. For information that is difficult to classify as either true or false, these fact-check outlets used their borderline fact-checking labels such as 'Mixture,' 'Unproven,'[2] 'Half-true,'[3] 'Center,'[4] 'Verdict pending'[5] and so on. These systems normally provide a verbal verdict with a complementary figure. Those figures, however, did not discriminate different natures among borderline results.

In this light, our study explores how different figures may help subjects perceive differences among borderline results. For this research purpose, our graphs visualize how different amounts of evidence are accumulated along the line of trueness (see Figure 1), i.e., the y-axis represents the amount of information. For each user, one out of six designs including None as the control condition was randomly assigned and shown. We did not specify the y-axis scales since each individual has a different baseline of the amount of information to make a judgment. We have conducted a user interview to understand how these graphs were perceived and could orient individual decisions.

In summary, we categorized borderline fact-checking messages into three conceptual types: high risk or divided, high uncertainty or lack of evidence, and neutral or mixed. If either risk- or uncertainty-aversion is in effect, the corresponding fact-checking message will act as a false signal rather than a neutral one to the audience, compared to mixed evidence. This work defined risk-aversion in the context of fact-checking by human discredit for risky borderline fact-checking conditions. Risk-aversion hypothesizes that people likely develop negative attitudes toward an online claim when exposed to a borderline fact-checking label with high risks, such as divided evidence, then when no label was shown to them. Uncertainty-aversion is defined to mimic discredit for uncertain words, wherein the context of fact-checking, this is represented as a lack of evidence. In our earlier work, we tested the hypothesis below:

- *H*1 **Post-stance of Borderline Messages:** Subjects are less willing to change their views towards claims tagged with borderline fact-checking conditions than those without any fact-checking. *(Partially Rejected)*

Subjects who were shown the Divided Evidence fact-check message did not develop substantially more negative post-stance towards a claim compared to those shown no fact-check messages; this indicates no support on the risk-aversion hypothesis. However, subjects who were shown the Lack of Evidence message developed considerably more negative post-stance towards a claim compared to those shown no messages; this indicates the presence of the uncertainty-aversion bias [33].

This paper utilizes the same data collected from this randomized survey and tests whether borderline fact-checking conditions differently affect users' intent to share an unproven claim across different platforms. Understanding this mechanism would benefit stakeholders on various online platforms whose final goal would be to mitigate fake news spreading via fact-checking. We do not impose a robust theoretical assumption but propose a testable hypothesis derived from an expectation that fact-checking results may yield desirable outcomes.

- *H*2 **Willingness to Share of Borderline Messages:** Subjects are less willing to share claims tagged with borderline fact-checking conditions than those without any fact-checking. *(Partially Rejected)*

2.3 Qualitative User Study

Augmenting the quantitative analysis, we newly conducted semi-supervised user interviews and present findings from qualitative analysis. This step is to complement the quantitative analysis and develop a deeper understanding of user perceptions. Specifically, we tried to answer the following question:

- Are subjects willing to share the claims tagged with the borderline fact-checking condition, and if so, why?

Based on semi-structured interviews with 10 participants, we learned the overall user experiences and hidden mechanisms that lead to their actions. We asked a set of questions about the perception of the fact-checking conditions they saw, the fact-checking conditions' effect on willingness to share on their social media, and how the fact-check conditions affect their change of stance on the rumor presented.

[2] https://www.snopes.com/fact-check-ratings.
[3] https://www.politifact.com/factchecks/list/?ruling=half-true.
[4] https://www.allsides.com/media-bias/center.
[5] https://www.washingtonpost.com/politics/2019/01/07/about-fact-checker.

2.4 Cognitive Bias for Opposite Fact-Check Condition to Pre-stance

Suppose risk- and uncertainty-aversion possibly distorts signals from borderline fact-checking results. Then, people's reactions to either true or false fact-checking results may be explained in the context of human biases involved in motivated reasoning, which potentially contributes to the belief and spreading of fake news [45]. One research claimed that motivated reasoning is especially persuasive when given information is vague, and the multiple logic is possible to interpret one situation [3].

Motivated reasoning can be divided into three different cognitive processes: prior attitude effect, confirmation bias, and disconfirmation bias [42]. An essential cognitive bias arises from the prior attitude effect; people who hold a strong attitude or belief tend to keep it rather than changing it to reflect any counter-evidence. This bias may sharply limit the effect of fact-checking since people stick to their prior stance, be it approval or disapproval, on a claim irrespective of how fact-checkers judge the claim. Our experiment will measure the strength of individuals' prior attitudes on a randomly assigned claim and examine on what extent the respondents keep their attitude despite a disconfirmation fact-checking condition.

The best-known bias among the three types of motivated reasoning is the confirmation bias. This bias leads people to selectively search and collect information congruent to their original beliefs and ignore any incongruent information [32], thereby creating filter bubbles of individuals on online social network [15]. Our experiments will observe disconfirmation bias that can directly limit the effect of fact-checking. People tend to scrutinize disconfirmatory information more thoroughly than confirmatory details. If confirmation bias helps people pay selective attention to supporting evidence, disconfirmation bias, though often roughly equated with confirmation bias, leads them to review counter-evidence critically and reject it [11]. A considerable fraction of online users is known to readily decide on political and social issues by their pre-existing thoughts and partisan identification, rather than constructing new narratives that better fit the interpretation [5]. Subsequently, we expect that a fact-checking label incongruent to one's prior view will activate cognitive skepticism or disconfirmation bias. We test:

- *H*3 **Disbelief-activated Disconfirmation Bias (hereafter Disapproval Bias):** Subjects with negative pre-stance toward unproved claims are less likely to correct their stance than those with positive pre-stance, by a disconfirming fact-checking result. *(Supported)*

3 METHODS & RESULTS

3.1 Experiments

3.1.1 Experimental Setup. We designed a randomized survey to test the effectiveness of fact-checking on a crowdsourcing platform [33]. The current study shares the same methodology as in this earlier work. It is based on a framework from existing literature that leveraged Amazon Mechanical Turk (MTurk) to perform experiments simulating an online news consumption environment [1]. We selected ten rumors listed below from the Unproven category on Snopes.com, selecting topics of interest to the general public to minimize suggestive prejudice. These rumors span various thematic categories, such as food & health, politics, life & entertainment to limit. Politically, rumors #4 and #5 are likely considered more trustworthy by conservatives, whereas #6 and #7 by liberals.

(1) Vegetarians live longer than meat-eaters.
(2) Drinking grape juice three times a day after being exposed to stomach flu will prevent you from sickness.
(3) Coffee serves as an effective mosquito repellent and protection against infection by the Zika virus.
(4) In 2017, Muslim immigrants committed 11,000 out of 13,000 total knife crime offenses in London.
(5) Policies that legalize recreational marijuana will worsen the opioid (drug) addiction epidemic in the U.S.
(6) A California bill would penalize police if they shot people carrying air guns or fake firearms.
(7) Withdrawing from the Iran nuclear deal will kill 100,000 Boeing jobs.
(8) A new study proves that men who marry chubby women are happier and live longer.
(9) Marilyn Monroe's IQ was measured at 168.
(10) Ariana Grande contacted the families of victims in the Manchester bombing terror attack and will pay for their funerals.

Participants were recruited through MTurk and were limited to young adults born in 1982-1999 in the U.S. A total of 11,145 workers (n=11,145, $M_{birthyear}$=1986, 52.93% female) enrolled the survey over a month period. Each subject is allowed to participate in the survey only once and has been monetarily rewarded for their participation. We collected survey respondents' personal information such as the birth year, gender, ethnicity, and political orientation. This research was reviewed and approved by the institutional review board (IRB) at the authors' institute.[6]

3.1.2 Experimental Study Design. Participants responded if they think a given claim is true or false based on their prior view (referred to as *pre-stance*). The pre-stance was rated on a 5-pt Likert scale (−2:definitely false, −1:false, 0:middle of the road, 1:true, 2:definitely true). In the survey, we have noted that 'not sure' respondents could choose the 'middle' option. Previous research found that a five-point scale is readily comprehensible to respondents and enables them to express their views [46]. Next, we presented participants with one of the six fact-checking decisions chosen randomly from our system: 1) None (no decision was given); 2) Mostly True; 3) Mostly False; 4) Mixed Evidence; 5) Divided Evidence, or 6) Lack of Evidence. To help participants understand the subtle differences in the meanings of the fact-checking decisions, we showed graphical representations in Figure 1.

Upon seeing a fact-check, respondents were asked to mark their willingness to share the content on a 4-pt Likert scale (1:definitely no; 2:no; 3:yes; 4:definitely yes). Then, we asked about the participants' stance toward the claim once again on a 5-pt scale (hereafter *post-stance*). For the control group users, they answered the same questions on willingness to share and post-stance in order without seeing any fact-checking labels. At the end of each survey, we debriefed the users that the tested service is a mock system, and all fact-checking decisions had been randomly generated.

[6] Approved by KAIST IRB. Approval number: KH2018-62.

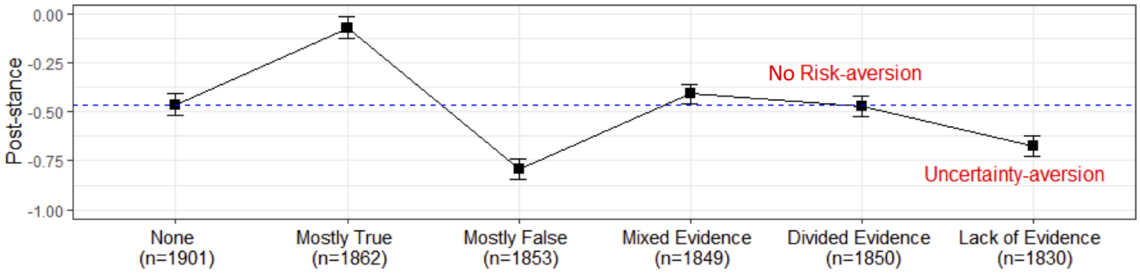


Figure 2: The effect of borderline fact-checking conditions, originally published in [33]. Risk-aversion is represented as the value difference between the post-stance of 'None' and 'Divided Evidence' conditions and uncertainty-aversion as the difference in post-stance of 'None' and 'Lack of Evidence' conditions. The error bars show the post-stance on rumors after seeing a hypothetical fact-checking condition (95% CI). The blue line shows the average post-stance of the None condition.

3.1.3 Basic Statistics on the Data Collected. Based on power analysis, we set up to collect at least 150 answers by six fact-checking conditions for each of ten unproven claims (i.e., the minimum required number of answers: 150 × 6 × 10 = 9K). The online experiment was conducted for one month as this was enough time to achieve this goal. Prior to analysis, incomplete responses or responses with duplicate IP addresses were discarded. Finally, a total of 11,145 valid users participated in this experiment. The distribution of the reported pre-stance was skewed to negative sentiment, i.e., participants reported on average that given claims are more likely false (negative=53%, neutral=20%, positive=27%, p<0.001). The pre-stance distributions were not equal among the ten rumors (χ^2=1928.32, df=36, p<0.001), indicating that some claims were less likely to be believed than others. We later discuss this discrepancy in detail in Section 4.2.

Table 1 shows the breakdown of the user counts based on their pre-stance and post-stance on the given ten claims; 27.8% and 5.3% of the respondents said the rumors are Definitely False and Definitely True, respectively, although all claims were neither true nor false at the time. Post-stance distribution is similar to that of pre-stance, yet fewer subjects take extreme views; 25.4% and 4.6% marked rumors as Definitely False and Definitely True.

Table 1: The distribution of users' pre-stance and post-stance upon seeing a hypothetical fact-checking message on unproven claims [33].

Post Pre	Def.False	False	Middle	True	Def.True	Total
Def.False	2,461	356	157	71	53	3,098
	(79.44%)	(11.49%)	(5.07%)	(2.29%)	(1.71%)	(27.80%)
False	232	2,083	366	175	10	2,866
	(8.09%)	(72.68%)	(12.77%)	(6.11%)	(0.35%)	(25.72%)
Middle	72	283	1,564	253	33	2,205
	(3.27%)	(12.83%)	(70.93%)	(11.47%)	(1.50%)	(19.78%)
True	54	197	445	1,605	83	2,384
	(2.27%)	(8.26%)	(18.67%)	(67.32%)	(3.48%)	(21.39%)
Def.True	14	28	60	156	334	592
	(2.36%)	(4.73%)	(10.14%)	(26.35%)	(56.42%)	(5.31%)
Total	2,833	2,947	2,592	2,260	513	11,145
	(25.42%)	(26.44%)	(23.26%)	(20.28%)	(4.60%)	

3.2 Result 1: Uncertainty-Aversion

Many commercial fact-checking systems employ borderline results that are neither true nor false. To answer the first research question, how do users perceive borderline fact-checking labels (e.g., Mixed, Divided, or Lack of Evidence, see Figure 1(b)), we examine the post-stance or rumors upon various fact-checking conditions.

Figure 2 depicts the average post-stance on rumors after being exposed to the random fact-checking messages. A horizontal axis includes all the six random conditions (no fact-checking result plus five results) with the number of subjects assigned to each condition in parenthesis. A vertical line marks the mean value and 95% confidence intervals (CI) of post-stance after each condition was given to subjects. Both pre-stance and post-stance were measured by a five-point scale as explained in Section 3.1.2.

The figure shows that survey users exposed to the Divided Evidence or Mixed Evidence condition show no significant difference to have seen no condition at all (i.e., None). This means that, according to our hypothesis, fact-checking conditions like Divided Evidence in our case that may induce risk-aversion behaviors did not prevent users from trusting a rumor, compared to presenting either a neutral result or no result at all (p>0.1 for both cases). In contrast, the Lack of Evidence condition displayed a far more negative post-stance than None (p<0.001), resulting in a post-stance that is even similar to have seen the Mostly False condition. The inability to collect sufficient evidence to verify a rumor may have led subjects to lean towards distrusting the rumor. These results confirm the working of uncertainty-aversion, but not risk-aversion.

We posit that such a difference among the three borderline conditions arise due to fact-checking message that further affects how people perceive information. It may be that people associate negative connotations to the terms 'lack of', but not to 'mixed.' Our experiment shows that different message choices can change people's post-stance towards identical rumors. As a result, fact-check decisions, depending on how they are presented to users, can bring unexpected outcomes, thereby need to be carefully devised.

3.3 Result 2: Willingness to Share

Figure 3(a) shows people's willingness to share upon the unproven claim. The x-axis represents the fact-checking conditions, and the y-axis the degree to which users would like to share the given claim. A higher value would indicate a stronger intent for sharing, as explained in Section 3.1.2. As values are aggregated across both the pre-stance and the four media type combinations, this figure shows a general trend. Overall, users are less willing to spread rumors when the fact-checking system shows the rumor is Mostly False than Mostly True. Again, when we control for the effects of pre-stance, the result still holds.

Among the borderline messages, Figure 3(b) shows that willingness to share increases for the Mixed Evidence and Divided Evidence conditions. There is however no meaningful difference between Lack of Evidence and None. Willingness to share is the highest for Divided Evidence. Later we perform user interviews to understand the latent mechanism that manifests these behaviors (i.e., *"Which borderline conditions do you think will most likely lead to content sharing?"*) and explain the outcomes in Section 4.3.

We further examine results across different platforms because people may find some platforms to disseminate unverified information than others. Hence, the type of platform was considered the confounding effect. Figure 3(c)–(f) shows the respondents' sharing intention across different platforms. Fact-checking condition showed a greater influence on instant messengers and emails than on social networking platforms; on Facebook or Twitter-like social media, respondents were not as reluctant to share the falsely labeled information as on instant messengers or emails (ANOVA test for willingness to share within Mostly True condition among four platforms: F(3,7444)=20.21, p<0.001, followed a posthoc analysis also confirms there is a significant difference between two groups (i.e., instant messenger–email and Facebook–Twitter) whereas no significant within each group).

3.4 Result 3: Disapproval Bias

Next, we investigate the change of users' stance according to the four fact-check conditions, None, Mostly True, Mostly False, and Mixed Evidence. To quantify the degree of stance change (*SC*), we transformed the 5-pt Likert scale to numeric ranges as Equation (1). The *SC* value of zero indicates that a user did not change their stance after the intervention. Likewise, a positive *SC* value indicates that a user has become more approving of the rumor after the intervention, whereas a negative indicates the opposite.

$$SC = post.stance - pre.stance. \tag{1}$$

Table 2 displays the result of ANOVA and Tukey's tests, which shows changes of stance difference between different intervention groups. The table shows the differences between the mean stance change between each pair of conditions. For instance, the mean stance change for the Mostly True group was 0.71, greater than the mean stance change for the Mostly False group (i.e., Mostly.True − Mostly.False = 0.71). Posthoc analysis reveals that most of the pair-wise differences are statistically significant.

The table indicates that stance change indeed happens in the direction guided by the fact-checking condition. The amounts that are denoted by the stance difference between None and Mostly True conditions, which are always negative. Between None and Mostly False, and between Mostly True and Mostly False are positive. In other words, stance change is observed in the direction of the intervention message for the conclusive messages (see Figure 1(a)). Among all stance change settings, we further paid attention to cases when fact-check conditions are opposite to those of the users' pre-stance sentiments. As a result, we observed *disapproval bias* by quantifying the difference between two pre-stances, positive and negative (see Equation (2) to be clear, where NP refers to negative pre-stance and PP to positive pre-stance):

$$f(\bar{SC}) = (|\bar{SC}_{None} - \bar{SC}_{False}| \in PP) - (|\bar{SC}_{None} - \bar{SC}_{True}| \in NP), \\ \text{if } f(\bar{SC}) \geq 0 \text{ then disapproval bias works.} \tag{2}$$

Relatively, the negative pre-stance group was less likely to change their beliefs (mean stance change difference between None and Mostly True is 0.59, p<0.001) than the positive pre-stance group (mean stance change difference between None and Mostly False is 0.82, p<0.001) as denoted in Table 2. We may conclude from this observation that a false negative error (i.e., stating that a rumor is false when it is true) may be harder to correct than a false positive (i.e., stating that a rumor is true when it actually is false) in combating false information online.

Table 2: ANOVA and post-hoc Tukey's honest significant difference (HSD) test results for Result 3. ΔSC refers to the magnitude difference of stance change between every pair of fact-checking conditions. Mixed Evidence was utilized as a representation of Borderline. We control the pre-stance by grouping them into three: All, Positive, and Negative. The two bold rows demonstrate the magnitude stance difference is considerably smaller when the pre-stance is negative, i.e., disapproval bias. This is explained in Equation (2).

Pre-stance	Group	ΔSC	p-value
All F(3,7461)=250.2, p<0.001	None − Mostly.True	-0.39	<0.001
	None − Mostly.False	0.32	<0.001
	None − Borderline	-0.03	0.62
	Mostly.True − Mostly.False	0.71	<0.001
	Mostly.True − Borderline	0.36	<0.001
	Mostly.False − Borderline	-0.35	<0.001
Positive F(3,1988)=136.9, p<0.001	None − Mostly.True	-0.04	0.78
	None − Mostly.False	**0.82**	**<0.001**
	None − Borderline	0.30	<0.001
	Mostly.True − Mostly.False	0.86	<0.001
	Mostly.True − Borderline	0.34	<0.001
	Mostly.False − Borderline	-0.52	<0.001
Negative F(3,3973)=157.8, p<0.001	**None − Mostly.True**	**-0.59**	**<0.001**
	None − Mostly.False	0.06	0.22
	None − Borderline	-0.23	<0.001
	Mostly.True − Mostly.False	0.66	<0.001
	Mostly.True − Borderline	0.37	<0.001
	Mostly.False − Borderline	-0.29	<0.001

To double-check if disapproval bias exists, we measure the strength of the pre-stance and post-stance signals in Figure 4(a)–(c). Users were grouped based on their pre-stance as positive (i.e., definitely true or true) or negative (i.e., definitely false or false) to control their pre-stance sentiment. As a baseline, these figures indicate that when users are exposed to fact-check conditions regardless of its messages, the users tend to regress to moderate views. This finding implies the positive role of fact-checking systems in that online

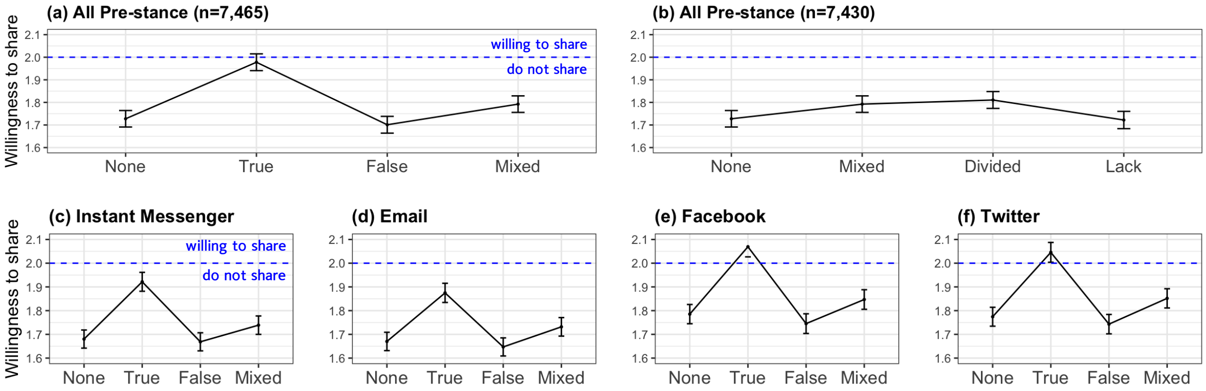


Figure 3: Results on willingness to share an unproven claim upon seeing a fact-checking message (95% CI). The y-axis indicates the willingness to share (1:definitely no; 2:no; 3:yes; 4: definitely yes), where the blue line represents the threshold value for a non-negative response. The x-axes compare the experimental conditions: None (or no condition), Mostly True, Mostly False, and Mixed Evidence. Borderline conditions are compared in (b). (a) shows that people are more likely to share rumors when they see positive fact-checks than negative fact-checks. (b) shows a willingness to share for various borderline labels, indicating a tendency on Lack of Evidence is different from Mixed Evidence and Divided Evidence. (c) to (f) show people's response to their willingness to share differs by the media platform; we observe an increased willingness to share unproven claims on social media, compared to on Messenger or Email.

users become less prone to maintaining extreme views when given a chance to reflect their judgment. Moreover, the average stance change for all conditions of users with positive pre-stance, depicted as a gap between two red lines in Figure 4(b), is greater than that of users with negative pre-stance, depicted as a gap in Figure 4(c), which also reveals the presence of disapproval bias.

3.4.1 Observing Disapproval Bias via Bayesian Inference. The presence of disapproval bias is noteworthy, as it may potentially limit the efficacy of any fact-checking service. Here, we provide an alternative view of this trend. We re-examine this trend by computing the odds ratio of the probability of the fact-checking conditions given by the stance change results by participants.

The detailed equations to compute the odds ratio are as follows, where SC is stance change, $cond$ is the fact-checking conditions to predict (Mostly True or Mostly False), SD is the same direction of the fact-check message from the pre-stance (i.e., control), OD is the opposite direction of the fact-check message (i.e., treatment), and OR is odds ratio:

$$\begin{gathered}
SC = pre.stance \times post.stance,\ \ where\ pre.stance \in \{-2, -1, 1, 2\}, \\
post.stance \in \{-2, -1, 0, 1, 2\},\ \ and\ SC \in \{-4, -2, -1, 0, 1, 2, 4\}, \\
p(cond|SC) = \frac{p(SC|cond) \times p(cond)}{p(SC)}, \\
Odds(SD) = \frac{p(cond_{SD}|SC)}{1 - p(cond_{SD}|SC)},\ Odds(OD) = \frac{p(cond_{OD}|SC)}{1 - p(cond_{OD}|SC)}, \\
OR = \frac{Odds(OD)}{Odds(SD)}.
\end{gathered} \tag{3}$$

The Type 1 cases in as displayed in Table 3 show, given that an online respondent initially judging a false claim as True but later changing views to False, the odds of the provided fact-checking condition being appropriate (i.e., Mostly False condition) is 11.83 times that of correction upon other kinds of interventions. Likewise, the Type 2 cases show, given that an online respondent initially judging a true claim as False but later changing views to True, the odds of the provided fact-checking condition being appropriate (i.e., Mostly True condition) is 10.97 times that of correction upon other kinds of interventions. As a result, correction rates in Type 1 and Type 2 are high, yet the rates of Type 1 are slightly higher than those of Type 2, which is also under our findings.

Table 3: The odds ratio (OR) and confidence interval (CI) of successful stance correcting cases: Type 1 representing the false positive error and Type 2, the false negative error.

	Given	Given	Predicted		
Type	pre-stance	post-stance	condition	OR	CI
1	Positive	Negative	Mostly False	11.83	3.34–25.28
2	Negative	Positive	Mostly True	10.97	1.80–17.05

4 STUDY VALIDATION

Having shown the biases involved in interpreting fact-checking messages, we now revisit the experiment assumptions to discuss potential limitations. We also present findings from small-group user interviews, which helps us better interpret the survey study.

4.1 Validity of the Data

In many cases, survey results are bound to multiple errors, and therefore we report some of the verification processes we took to ensure the validity of results. The Chi-squared test of independence identified that fact-checking conditions were randomly distributed across the ten rumor claims. No significant interaction was

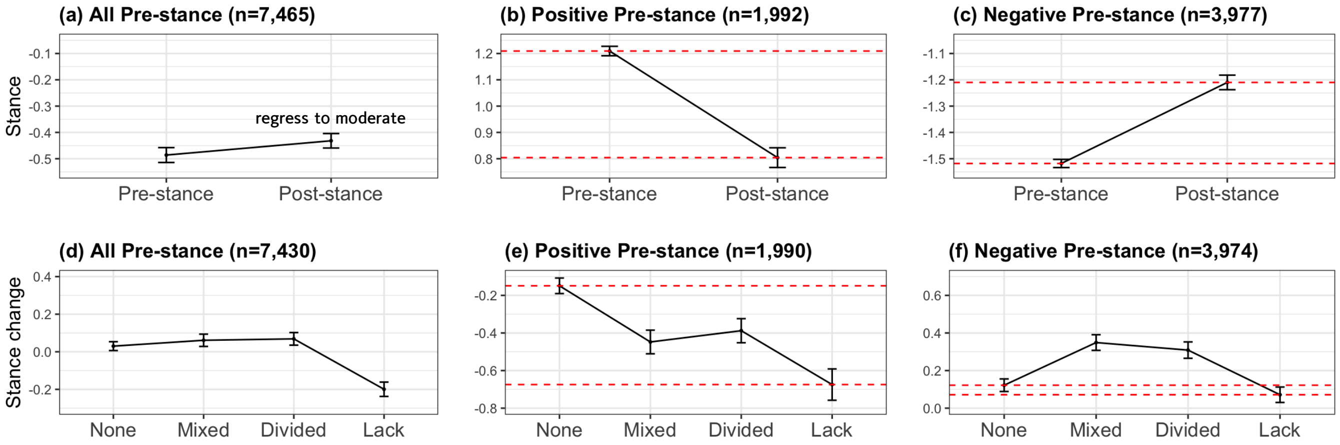


Figure 4: Stance change after being exposed to a fact-checking condition (95% CI). The x-axis in (a)–(c) compares users' stance before and after seeing the fact-check messages while the y-axis indicates the degree of users' stance, where the range is from -2 to 2. The x-axis in (d)–(f) compares the experimental conditions: No condition, Mostly True, Mostly False, and Mixed Evidence while the y-axis indicates the magnitude of stance change in Equation (1). (a–c) The average sentiment of post-stance moves to the opposite direction irrespective of the pre-stance sentiment, which may imply that fact-checking helps people re-evaluate their judgment and take less radical attitudes. The gap between the two red lines on (b) and (c) represent the mean difference between pre-stance and post-stance (i.e., average stance change), and the greater gap on (b) manifests that disapproval bias exists. (d–f) Stance change for different borderline fact-checking labels, indicating a tendency on Mixed Evidence and Divided Evidence is different from on Lack of Evidence. The gap between the two red lines on (e) and (f) represents the mean difference between two fact-checking conditions; one is None condition and another is Lack of Evidence (i.e., uncertainty-aversion): notably, disapproval bias is shown not only at the Mixed Evidence case but also at Lack of Evidence.

observed between the assigned conditions and claims (χ^2=44.21, df=45, p=0.51), and one-way ANOVA for pre-stance did not find any significant difference across the six states (F=0.77, p=0.57). We also checked what fraction of the respondents always picked the first (or the left-most) choices throughout the survey. This resulted in 0.35% (39/11,145) of all subjects, which is a negligible proportion.

When choosing the ten claims, we tried to find topics of general interest. However, some topics may be more intuitive to understand than others, causing external validity problems. The users' total response time was observed to identify any bias in perception difficulty across the claims and to check whether users took a reasonable amount of time on the survey. The total distribution of time spent on the survey is skewed. Most turkers completed their tasks within 10 minutes (98.80%) and the median time taken was 96 seconds, but a small fraction took a disproportionately large amount of time (e.g., several hours). Nonetheless, 99 percent completed the task within 12.2 minutes. Examining those users who took significantly long sessions, we find no correlation between the time spent and claims (χ^2=8.08, df=9, p=0.53).

4.2 Validity of the Experiment

The left two graphs of Figure 5 present the percentage of negative pre-stance over ten claims and its association with that of positive pre-stance, and the right-most graph indicates the relationship between the proportion of negative pre-stance and degree of disapproval bias per claim. The perceived falseness or trueness could differ across 'unproven' claims, as shown in the left-most graph. Simultaneously, the middle graph shows that 8 of 10 unproven claims showed more negative pre-stance than the positive one on average. Given this predominant negative pre-stance, one can suspect that the larger stance change for positive pre-stance in Figure 4(b) than for negative one in Figure 4(c) might not necessarily come from disapproval bias but from subjects' propensity toward the majority opinion (e.g., some subjects may access external resources, so they may acknowledge the mainstream stances). Subjects with a positive pre-stance may be more willing to change to the opposite position than those with a negative one, knowing that others tend to believe the opposite. If this were the case, claims with stronger negative pre-stance would show stronger disapproval bias.

Figure 5(c), however, does not support this hypothetical case. It does not show a clear relationship between the proportion of negative pre-stance and the degree of disapproval bias across the claims. If any, a mild, negative association (cor=-0.53, p=0.11), suggests a weaker disapproval bias for a claim with stronger negative pre-stance. This association suggests that if we replicate the current experiment with unproven claims of a more balanced pre-stance between true and false, we are likely to observe a more substantial disapproval bias rather than a weaker one. To conclude, we assess with caution that the disapproval bias observed in this study is not a byproduct of predominant negative pre-stance.

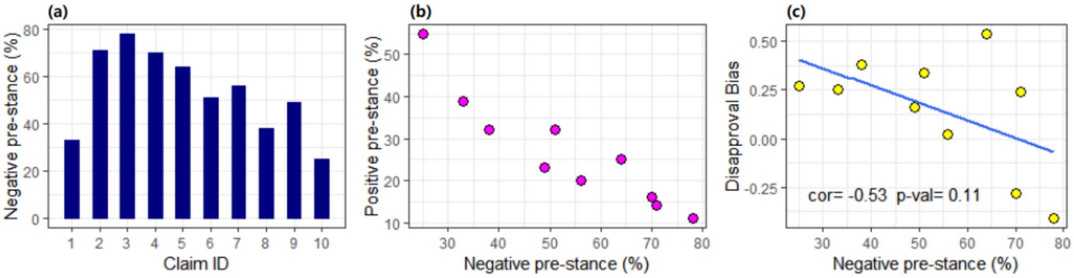


Figure 5: The variance across the ten claims shown by the ratio of negative pre-stance, distribution of negative and positive pre-stance, and the level of disapproval bias, i.e., $f(\bar{SC})$ in Equation (2), as a function of negative pre-stance level. Although all were from the 'unproven' category from the source data, this figure demonstrates that some claims could be perceived as more false than others. The resulting bias, as a result, showed a slight negative tendency (cor=-0.53, p=0.11).

4.3 User Interview

4.3.1 ***Participants and Procedure.*** We conducted semi-structured interviews with ten individuals to understand their overall experiences and the latent mechanisms that drive their action tendencies. The median age was 25 years old, and 40% were females. Users responded that they were exposed to the news 3.5 times a week on average (66.7%) via online news aggregation sites.

The interview included a set of questions about the fact-checking conditions, presented along with the survey items, and the fact-checking conditions' effect on willingness to share on social media. The questions were as follows:

(1) *Perception*: What do you think is the meaning of the fact-checking condition you shown?
(2) *Willingness to share*: Which fact-check condition will most likely lead to content sharing and why?
(3) *Stance changes*: Did the fact-check condition affect your change of stance on the rumor presented?

All user interviews were audio-recorded and transcribed. The two authors independently coded the user transcriptions and iteratively discussed the identified codes until they reached a consensus.

4.3.2 ***Results.*** Most of the interviewees well understood fact-checking conditions as we intended. Besides, eight out of ten said they were influenced by fact-check conditions, regardless of the type of fact-check condition they provided. P2, P3, P4, and P5 said that even though the given borderline fact-check condition did not help determine the claim's veracity, the information represented by the borderline label helped them decide whether to trust the rumor or not. For the borderline fact-check labels, P4 said:

> *"The borderline label shown did not provide any clear verdict on whether the claim is true or false, but it certainly made me think carefully about the claim again."*

The Mostly True condition was selected as the fact-check condition that users are most likely to share on their social media (n=4), followed by Mostly False (n=3) and Divided Evidence (n=3). It is also intriguing that a few interviewees (n=3) also responded that they would share the unproven claims more if Divided Evidence is given as a fact-checking result. This finding aligns well with Figure 3(b); Seven out of ten users responded that they are most likely to share the claim of the Divided Evidence condition on their social media among the three borderline fact-check conditions.

P1, P4, P5, and P9 said they wanted to share the claim of Divided Evidence on social media to initiate the discussion with others. P1, P4 mentioned that sharing a controversial claim on social media could get more attention from social media friends than sharing the truth. P7, P8 explained that fun and interest are one of the most crucial factors in sharing information. Mainly, there is one intriguing comment from P7 why the respondent is willing to share the misinformation even after being noticed that the given information is Mostly False.

> *"Fun is a critical criterion for sharing something online for me. So I would share information or misinformation only if it is fun and not hostile to people to be sent."*

We also noticed that some interviewees (P3, P4, P5) posit Mostly False and Lack of Evidence to be similar to falseness. This aligns with our findings on Uncertainty-aversion presented in Result 1.

Regarding the fact-check label design, some users (P1, P2, P6, P8, and P9) wished the graph would display more additional information, such as the number of shares or likes. P6, P8, and P9 mentioned that the graph's shape alone was not enough to effectively convey the meaning of the fact-check label. P1 and P2 thought that the number of likes for a claim represents the degree of the claim's influence power. P5 and P10 said that the data source needs to be added to the graphs with the fact-check labels in order to provide more reliable information.

5 DISCUSSION

Our findings are two-fold: 1) uncertainty-aversion may unintendedly nurture a negative stance toward an unproven claim due to lack of information, and 2) false suspicion may be harder to correct than a false belief by a fact-checking result due to its more substantial disconfirmation bias. We discuss these points in detail.

5.1 The Role of Borderline Fact-checks

According to the decision-making theory, risk-aversion means the human distaste for risky situations in terms of subjective utility, which has been widely observed [8, 36]. We found a strong proof of uncertainty-aversion, yet no grounds for risk-aversion behavior for fact-checking conditions (see Figure 2 for exploring risk- and uncertainty-aversion cases, respectively).

The presence of uncertainty-aversion has practical implications for fact-checking business. Claims with temporarily insufficient information are bound to increase as more users start relying on fact-checking systems. Their fact-checking speed will ultimately lag behind the claim-registering speed. As shown in our experiments, 'lack of information' will not be considered neutral but raise negative post-stance to users. The interviews also revealed that users are more likely to share controversial claims (e.g., Divided Evidence) to get more attention from others. This was described as "fun" or "novel information" by participants. In short, the more popular a fact-checking system becomes, the more false-negative errors will likely be yielded. These increasing errors have nothing to do with the inherent accuracy of the fact-checking system. However, system-makers may minimize this problem by developing a better label that does not strongly convey the possible negative association with the term 'lack of.' For instance, they may experiment on how results such as 'lack of sufficient time' and 'in preparation' cause different responses from Lack of Evidence. The implication is not limited to linguistic expressions such that system-makers can include graphical illustrations. Therefore, they should also be aware of unintended consequences of increased uncertainty in fact-checking messages.

The implication above can be adapted to the willingness to share. We acknowledge that Divided Evidence can increase people's willingness to share the unproven claims (i.e., no risk-aversion), whereas Lack of Evidence can decrease willingness (i.e., uncertainty-aversion works). In this light, we need to carefully design the fact-checking systems not unintentionally to expedite spreading rumors. One recent qualitative study explored that users on social media now pay attention to online misinformation and are willing to pay attention to education or design solutions [18].

5.2 The Role of Pre-Stance on Misinformation

This research also found a consistent pattern that fact-checking effects in correcting views and mitigating the spread of falsehoods are influenced by individuals' initial stance toward the claims. In particular, Result 2 demonstrated that fact-checking interventions play the desired role of correcting users' beliefs only in certain circumstances. Those who initially disapproved of a claim are less likely to change their views upon exposure to interventions than those who favorably view the claim. We call this tendency *disapproval bias,* which represents that people activate a stronger form of disconfirmation bias when starting with disbelief than a belief. As a consequence of this bias, a false-negative error (i.e., stating that a rumor is false when it is true) is more challenging to correct than a false-positive error (i.e., stating that a rumor is true when actually false) in combating falsehoods online. If the proportion of disinformation — or purposely manufactured false news — is increasing online, people's overall cynicism and negative stance on online claims will also increase. As a consequence, the effort of fact-checking to verify correct information will have to battle with disapproval bias increasingly. How to minimize such bias should become a more immediate task soon.

Results in this study reveal motivated reasoning at play in people's responses to fact-checking. Prior attitude effect is represented by the participants who maintain their stance irrespective of fact-checking interventions. 61.43% (895 number) of the subjects confirmed that prior attitude effect is the primary form of motivated reasoning. Next is represented by the participants who correct their stance towards the opposite direction upon an intervention. They make up 38.57% (562 number) of the conditioned participants. The most intriguing cases arise when considering the following two groups: participants who changed the stance and the participants who stuck to the initial position when the fact-checking result opposite from pre-stance is given. Our study is the first to report how disconfirmation bias interacts with pre-stance — i.e., individuals with negative pre-stance were less willing to shift their views than those with positive, which we called disapproval bias. What processes, then, lead to this asymmetry? We hypothesize that a critical review already shapes negative pre-stance on a claim and, therefore, is less likely to be overturned upon fact-checking.

Psychological theories related to cognitive dissonance may explain this gap. According to the theory, mental discomfort is triggered by a situation in which a person's belief clashes with new evidence provided to the person. Confronting with facts that contradict personal beliefs, people will find a way to resolve the contradiction to reduce their discomfort [14]. Given the nature of online claims, it likely takes less cognitive effort to reject them or take a neutral stance than to approve them. Hence, the magnitude of dissonance for the negative pre-stance group will be more extensive; it likely requires more cognitive efforts to 'approve' what was rejected initially, which may have resulted in reduced correction rates for these participants. The last mechanism is represented by the participants who shifted views yet in an unexpected direction. They make up the smallest proportion of all participants (6.58%). While surprising, this may be another function of fact-checking, merely providing an opportunity to reconsider one's thoughts.

5.3 More Analyses based on Demographics

Regarding the political orientation of participants, we repeated analysis by political leaning: Liberal (N=3,758), Middle (N=3,761), and Conservative (N=2,008) [33]. Based on the ANOVA followed by Post-hoc Tukey's HSD tests, all three subgroups showed uncertainty-aversion (Liberal: the post-stance difference between None and Lack of Evidence is 0.18, P=.08, Middle: 0.25, p<0.01, Conservative: 0.28, P=0.03). No subgroups showed risk-aversion. This reinforces our findings. We also found conservatives to show a more considerable negative stance change upon seeing the Lack of Evidence message than others. This may imply that conservative young adults showed a higher degree of uncertainty-aversion bias. This finding aligns with recent research that showed conservatives are more likely to respond to negative information online [21].

In addition, considering participants' backgrounds, we have collected the ethnic information of the respondents and checked whether the same perception bias (i.e., uncertainty-aversion) would be found for three major ethnic subgroups, Asian (N=1,754), Black (N=955), and White (N=7,544) for all ten claims. Based on the ANOVA followed by Post-hoc Tukey's honest significant difference tests, only White shows uncertainty-aversion (post-stance difference between None and Lack of Evidence is 0.30, p<0.001) while all subgroups show no risk-aversion. As it may be too early to conclude why only the White subgroup shows significant uncertainty aversion bias, e.g., it may be due to the different sample sizes by subgroups, we leave this question to future work.

Regarding the willingness-to-share by topical categories, the ANOVA analysis of the limited ten claim set showed that political topics were less likely to be shared than our participants' food/health/entertainment topics. However, we acknowledge this is an early finding and will keep the idea for future direction.

5.4 Implication and Future Direction

The perception, disapproval bias, is especially crucial amid COVID-19 in practice. For instance, as more people once refuse vaccinations based on the current anti-vaccination claims [41], it would be hard to change their thoughts, although strong fact-checking evidence debunking the claims are given due to disapproval bias. One recent study also found that intensified susceptibility to misinformation is connected to reduce specific preventive measures like mask-wearing and social distancing [37]. In the current pandemic, one recent work developed an Awareness index by calculating the participants' knowledge degree then revealed that people mainly ignore medical news regardless of their educational histories. Therefore, much effort is needed on fact-checking COVID-19 in social media [16]. The fake news phenomena have led to legislative initiatives for handling uncensored content on social media and news platforms; Germany, France, and Ireland have passed or considered legislation that penalizes fake news distribution.

Concerning other borderline conditions such as Divided and Lack of Evidence, Figure 4(d) depicts that stance change by Lack of Evidence is significantly more negative than stance changes by other neutral or borderline results. This observation is consistent with or redundant primarily to what we found in the first result (i.e., uncertainty-aversion). We also excluded the Lack of Evidence condition on the second analysis (i.e., disapproval bias) to remove any confounding factors such as uncertainty-aversion that might affect the target mechanism. In the meantime, we find disapproval bias for the Lack of Evidence condition as shown in Figure 4(e) and (f): this disapproval bias value (i.e., the gap on (e) – the gap on (f)) was 0.48, which is about two times greater than that of the second result, which was 0.23 (can be calculated at Table 2). Thus, there may be interactions between uncertainty-aversion and disapproval bias, and we can further explore the interactions in future studies.

We demonstrate that the design choice may trigger unexpected perception biases in the users for the intervention design on fact-checking systems. For example, some users may perceive Lack of Evidence as an insubstantial claim, although we did not investigate further. Future studies can examine more effective fact-checking representations on texts and graphics in greater detail and explore the relationships between the identified biases and various other kinds of borderline conditions.

6 CONCLUDING REMARK

This work identified an unexpected perception bias caused by borderline fact-checking messages and their initial stance toward claims. This work also contributes to understanding how borderline fact-checking messages affect users' content sharing and bring insights on real user perceptions via a user interview.

Misinformation critically affects political campaigns [31, 44] and other domains that we explore. This current study collected answers from 11K users over a broad set of real-world misinformation. We also provide a qualitative explanation for the randomized survey's quantitative findings by conducting semi-supervised user interviews to learn the user experience and perception of different fact-checking conditions. Hence, the data and findings add value to the literature. Many of the claims these days belong to the grey area are neither true nor false, but somewhat nuanced and partially accurate. No literature, to our knowledge, has examined the effect of borderline fact-checking messages. Hence, our findings on the unexpected perception bias towards those messages are novel and would be of interest to relevant practitioners.

The negative-skewed propensity, the first finding of the current work, can be more dangerous when it meets disapproval bias, the second finding. Disapproval bias hampers users to change their stances once they initially get negative perceptions towards given information. Considering disapproval bias, shown in the current study, it is also feasible to provide rich evidence towards firm unbelievers explaining why the given claim is the fact. On the other hand, in a drastic case, fact-checking outlets may reject ratings altogether and introduce users to various aspects of online claims to help them make independent decisions. Full Fact, the UK-based fact-checking charity, adopts this strategy. Its founder Will Moy notes that moving away from rating claims will allow a 'less combative, more collaborative approach to fact-checking [19].'

Fact-checking is promptly becoming a solution to tackle false information online. The current study based on large-scale survey experiments offers insights on the effects and limits of fact-checking. Future studies can build upon our work to further explore how different factors jointly determine fact-checking effectiveness.

ACKNOWLEDGMENTS

This work was supported by the Institute for Basic Science (IBS-R029-C2) and the Basic Science Research Program through the National Research Foundation funded by the Ministry of Science and ICT in Korea (No. NRF-2017R1E1A1A01076400).

REFERENCES

[1] Mahmoudreza Babaei, Abhijnan Chakraborty, Juhi Kulshrestha, Elissa M Redmiles, Meeyoung Cha, and Krishna P Gummadi. 2019. Analyzing Biases in Perception of Truth in News Stories and Their Implications for Fact Checking.. In *proc. of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 139.

[2] Meital Balmas. 2014. When Fake News Becomes Real: Combined Exposure to Multiple News Sources and Political Attitudes of Inefficacy, Alienation, and Cynicism. *Communication Research* 41 (2014), 430–454.

[3] David M Bersoff. 1999. Why Good People Sometimes Do Bad Things: Motivated Reasoning and Unethical Behavior. *Personality and Social Psychology Bulletin* 25, 1 (1999), 28–39.

[4] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2, 1 (2011), 1–8.

[5] Toby Bolsen, James N Druckman, and Fay Lomax Cook. 2014. The Influence of Partisan Motivated Reasoning on Public Opinion. *Political Behavior* 36 (2014), 235–262.

[6] 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), e2020043118.

[7] Grégoire Burel, Tracie Farrell, Martino Mensio, Prashant Khare, and Harith Alani. 2020. Co-spread of Misinformation and Fact-Checking Content During the COVID-19 Pandemic. In *proc. of the International Conference on Social Informatics (SocInfo)*. 28–42.

[8] Colin Camerer and Martin Weber. 1992. Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty* 5, 4 (1992), 325–370.

[9] 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 (2020), 1–23.

[10] Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. In *proc. of the Association for Information Science and Technology (ASIS&T)*, Vol. 52. 1–4.

[11] Kari Edwards and Edward E Smith. 1996. A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology* 71, 1 (1996), 5–24.

[12] Larry G Epstein. 1999. A Definition of Uncertainty Aversion. *The Review of Economic Studies* 66, 3 (1999), 579–608.

[13] Jessica T. Feezell, John K. Wagnera, and Meredith Conroy. 2021. Exploring the effects of algorithm-driven news sources on political behavior and polarization. *Computers in Human Behavior* 116 (2021), 106626.

[14] L Festinger. 1962. *Cognitive Dissonance*. Scientific American. https://www.scientificamerican.com/article/cognitive-dissonance/. Accessed: 2021-02-14.

[15] Seth Flaxman, Sharad Goel, and Justin M Rao. 2016. Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly* 80, S1 (2016), 298–320.

[16] Marco Furini, Silvia Mirri, Manuela Montangero, and Catia Prandi. 2020. Untangling between fake-news and truth in social media to understand the Covid-19 Coronavirus. In *IEEE Symposium on Computers and Communications (ISCC)*. 1–6.

[17] 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. In *proc. of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, Vol. 2.

[18] Christine Geeng, Savanna Yee, and Franziska Roesner. 2020. Fake News on Facebook and Twitter: Investigating How People (Don't) Investigate. In *proc. of the ACM CHI Conference on Human Factors in Computing Systems (CHI)*. 1–14.

[19] Lucas Graves and Federica Cherubini. 2016. *The rise of fact-checking in Europe*. Reuters Institute. https://tinyurl.com/1h0ny63x. Accessed: 2021-02-14.

[20] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *proc. of the Web Conference (WWW)*. 729–736.

[21] Jiyoung Han, Meeyoung Cha, and Wonjae Lee. 2020. Anger contributes to the spread of COVID-19 misinformation. *Harvard Kennedy School (HKS) Misinformation Review* 1 (2020). https://doi.org/10.37016/mr-2020-39

[22] Aniko Hannak, Drew Margolin, Brian Keegan, and Ingmar Weber. 2014. Get Back! You Don't Know Me Like That: The Social Mediation of Fact Checking Interventions in Twitter Conversations.. In *proc. of International AAAI Conference on Weblogs and Social Media (ICWSM)*. 187–196.

[23] Shan Jiang, Simon Baumgartner, Abe Ittycheriah, and Cong Yu. 2020. Factoring fact-checks: Structured information extraction from fact-checking articles. In *proc. of the Web Conference (WWW)*. 1592–1603.

[24] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2018. Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation. In *proc. of the ACM International Conference on Web Search and Data Mining (WSDM)*. 324–332.

[25] Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. In *proc. of the Web Conference (WWW)*. 591–602.

[26] Sejeong Kwon, Meeyoung Cha, and Kyomin Jung. 2017. Rumor Detection over Varying Time Windows. *PloS one* 12, 1 (2017), e0168344. https://doi.org/10.1371/journal.pone

[27] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[28] Kalev Leetaru. 2017. *The Backfire Effect And Why Facebook's 'Fake News' Warning Gets It All Wrong*. Forbes. https://bit.ly/3dwwxJh. Accessed: 2021-02-14.

[29] Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. 2015. Real-time Rumor Debunking on Twitter. In *proc. of ACM Conference on Information and Knowledge Management (CIKM)*. 1867–1870.

[30] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting Rumors from Microblogs with Recurrent Neural Networks. In *proc. of the International Joint Conferences on Artificial Intelligence (IJCAI)*. 3818–3824.

[31] Drew B Margolin, Aniko Hannak, and Ingmar Weber. 2018. Political Fact-Checking on Twitter: When Do Corrections Have an Effect? *Political Communication* 35, 2 (2018), 196–219.

[32] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175.

[33] Sungkyu Park, Jaimie Yejean Park, Jeong-han Kang, and Meeyoung Cha. 2021. The presence of unexpected biases in online fact-checking. *Harvard Kennedy School (HKS) Misinformation Review* (2021). https://doi.org/10.37016/mr-2020-53

[34] 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.

[35] Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying Misinformation in Microblogs. In *proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1589–1599.

[36] Matihew Rabin. 2013. Risk aversion and expected-utility theory: A calibration theorem. In *Handbook of the Fundamentals of Financial Decision Making: Part I*. World Scientific, 241–252.

[37] Jon Roozenbeek, Claudia R. Schneider, Sarah Dryhurst, John Kerr, Alexandra L. J. Freeman, Gabriel Recchia, Anne Marthe van der Bles, and Sander van der Linden. 2020. Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science* 7, 201199 (2020).

[38] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A Hybrid Deep Model for Fake News Detection. In *proc. of ACM Conference on Information and Knowledge Management (CIKM)*. 797–806.

[39] 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.

[40] Jeff Smith. 2017. *Designing Against Misinformation*. Medium.com. https://tinyurl.com/11vo86wb. Accessed: 2021-02-14.

[41] Dominik Andrzej Stecula, Ozan Kuru, and Kathleen Hall Jamieson. 2020. How trust in experts and media use affect acceptance of common anti-vaccination claims. *Harvard Kennedy School (HKS) Misinformation Review* 1, 1 (2020). https://doi.org/10.37016/mr-2020-007

[42] Charles S Taber and Milton Lodge. 2006. Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science* 50, 3 (2006), 755–769.

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

[44] Nathan Walter, Jonathan Cohen, R Lance Holbert, and Yasmin Morag. 2020. Fact-checking: A meta-analysis of what works and for whom. *Political Communication* 37, 3 (2020), 350–375.

[45] Brian E Weeks and R Kelly Garrett. 2014. Electoral consequences of political rumors: Motivated reasoning, candidate rumors, and vote choice during the 2008 US presidential election. *International Journal of Public Opinion Research* 26, 4 (2014), 401–422.

[46] Robert M Worcester and John Downham. 1986. *Consumer market research handbook*. Sole distributors for the USA and Canada, Elsevier Science Pub. Co.

[47] Dustin Wright and Isabelle Augenstein. 2020. Claim Check-Worthiness Detection as Positive Unlabelled Learning. *Findings of EMNLP. Association for Computational Linguistics* (2020).