Harvard Kennedy School Misinformation Review[1]
January 2021, Volume 2, Issue 1

Reprints and permissions: misinforeview@hks.harvard.edu
DOI: https://doi.org/10.37016/mr-2020-53
Website: misinforeview.hks.harvard.edu

Research Article

The presence of unexpected biases in online fact-checking

The increasing amount of information online makes it challenging to judge what to believe or discredit. Fact-checking unverified claims shared on platforms, like social media, can play a critical role in correcting misbeliefs. The current study demonstrates how the effect of fact-checking can vary by several factors. We show that fact-checking helps self-correct one's views among young adults. However, this effect is weaker for individuals who perceived the claim negatively at first. Furthermore, borderline messages like "Lack of Evidence" can be perceived as false rather than neutral. We explain these biases via human cognitive mechanisms that avoid risk and uncertainty.

Authors: Sungkyu Park (1), Jaimie Yejean Park (2), Jeong-han Kang (3), Meeyoung Cha (1,4)
Affiliations: (1) Data Science Group, Institute for Basic Science (IBS), South Korea, (2) Mobile Communications Division, Samsung Electronics, South Korea, (3) Department of Sociology, Yonsei University, South Korea, (4) School of Computing, Korea Advanced Institute of Science and Technology (KAIST), South Korea
How to cite: Park, S., Park, J. Y., Kang, J., & Cha, M. (2021). The presence of unexpected biases in online fact-checking. *Harvard Kennedy School (HKS) Misinformation Review*, *2*(1).
Received: September 9th, 2020. Accepted: January 12th, 2021. Published: January 27th, 2021.

Research questions

- [Self-correction Role] What effect does fact-checking have on individuals who take an extremely positive or negative initial view toward the claim?
- [Perception Bias] Are borderline fact-checking messages such as Mixed Evidence, Divided Evidence, and Lack of Evidence perceived as neutral?
- [Perception Bias] What are the cognitive mechanisms that trigger people's *reactions* when they see borderline fact-checking messages or opposing messages to their initial views?

Essay summary

- Fact-checking plays a critical role in fighting misbeliefs online. This research reveals the unexpected and diminished effect of fact-checking due to perception biases.
- Fact-checking platforms are bound to host many unverified claims that are neither entirely true nor false and inevitably produce borderline tags to those claims. We show the effect of such borderline messages by examining how readers perceive them.
- We experimented with 11,145 young adults and observed changes in their views on 10 real-world unproven claims after presenting a hypothetical fact-checked message.

[1] *A publication of the Shorenstein Center for Media, Politics and Public Policy, at Harvard University, John F. Kennedy School of Government.*

- Claims marked as “Lack of Evidence” were perceived similarly as false information, unlike other borderline tags such as “Mixed Evidence” or “Divided Evidence.” This is related to an uncertainty-aversion response due to insufficient information.
- Disconfirmation bias or reluctance to change views was observed when users saw a definite yet opposing tag like “Mostly True” to a claim they initially perceived negatively.
- While fact-checking is known to be effective in relaxing extreme views, the presence of various human cognitive biases indicates that fact-checking does not always produce the desired effect and that the self-correction effect varies by the wording in fact-checking messages.

Implications

Misinformation and fact-checking

Online misinformation and fake news have become a grand challenge over recent years, where a vast majority of the world's population relies on the Internet for finding information. As more people utilize online news media and social networks to discover news, the danger of encountering false news information increases (Kwon et al., 2013). This can have a detrimental impact, for instance, on political and financial decisions (Vosoughi, Roy, & Aral, 2018; Stecula et al., 2020). For example, studies have explored relationships between misinformation and the COVID-19 pandemic (Hall & Albarracín, 2020; Zarocostas, 2020). Based on large-scale surveys, one study found that people using social media as a significant information source tend to believe in online conspiracies and do not comply well with public health guidelines (Allington et al., 2020).

In this light, fact-checking is considered a significant and viable solution for combating online fake news. Fact-checking systems continually monitor media and political sources, determine claims to be checked based on prioritization, then review claims against existing fact-checks or authoritative sources before finally flagging fake news content and providing contexts (Hassan et al., 2017). Hereafter, we use the term 'system' when this process can be implemented as one of the modules inside the social media platforms to provide fact-checking tags. In contrast, we use 'outlet' when the monitoring is operated independently from the platforms. Numerous independent fact-checking outlets have emerged, including Snopes, Politifact, FactStream, FactCheck, StopFake, and The Conversation, and so have in-house services for global platforms like Google and Facebook. The desired effects of online fact-checking systems are twofold: 1) helping users come to an informed viewpoint and an accurate understanding of a claim; and 2) stopping the spread of false claims. The current work shows experimental results that reveal fact-checking messages lead participants to choose less extreme views, irrespective of the message type.

Borderline messages in fact-checking

However, a critical yet practical question has been overlooked, which is "what if fact-checking systems cannot classify a claim as either true or false?" Because many claims do not lend themselves to simple true-or-false verdicts, many fact-checking outlets adopt multi-scale judgments. Furthermore, in its attempt to minimize both false-positive and false-negative errors, an accurate classifier will inevitably hold a decision or yield to many indecisive, borderline choices that may not be interpreted as a neutral signal to the online audience. The effect of such unintended consequences may even diminish the future credibility of fact-checking systems. Even tagging a general warning could decrease the perceived accuracy of correct information online (Clayton et al., 2020).

In estimating the effect of borderline messages, we need to distinguish two similar yet separate cognitive burdens in decision-making: risk and uncertainty. We hypothesize that risk and uncertainty may trigger cognitive burdens and, subsequently, people tend to avoid those burdens by discrediting information associated with risk and uncertainty.

Risk represents the probability of an event or an outcome of interest, such as a failure (Epstein, 2004). For example, a high-risk investment suggests that both massive success and colossal failure are highly probable. If we apply this definition to fact-checking, a borderline decision is defined to be associated with 'high risk' if and only if both true and false verdicts are highly probable simultaneously. Betting on either true or false entails a high risk of failure when fact-checkers' decisions on a claim are sharply divided between Definitely True and Definitely False, i.e., the Divided Evidence case in Figure 1.

In comparison, *uncertainty* represents a void of information that helps formulate risk in probability (Epstein, 2004). Under uncertainty, the risk is not measurable and, therefore, the expected magnitude of failure or success is ambiguous. A fact-checked decision is associated with 'high uncertainty' if the system fails to collect sufficient information, for instance, when fact-checkers are unfamiliar with a given claim and are reluctant to conclude it ('Lack of Evidence' in Figure 1). Finally, a fact-checked decision is deemed truly *neutral* if it is neither associated with high risk nor high uncertainty, for example, when fact-checkers exhibit a full spectrum of choices from Definitely True to Definitely False on a single claim ('Mixed Evidence').

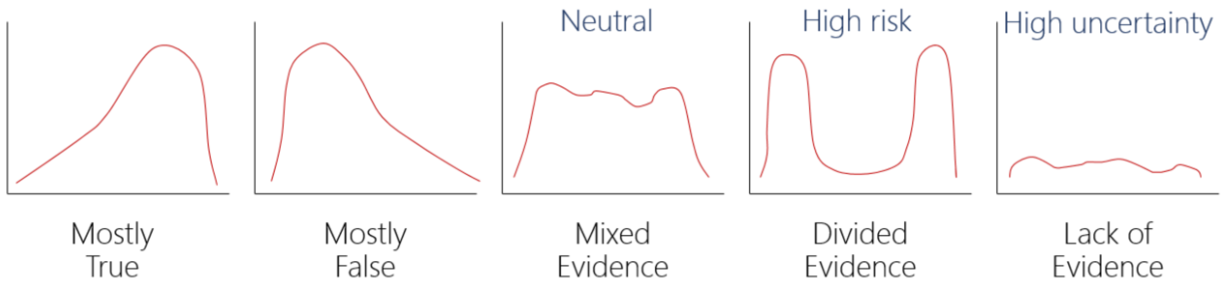


Figure 1. Designs of five hypothetical fact-checking messages. *The three plots on the right are borderline messages. Mixed Evidence means when fact-checkers gave the full spectrum of true to false decisions, Divided Evidence means when fact-checkers are sharply divided in their decisions, and Lack of Evidence represents when there is not enough information for judgment. The graph title descriptions marked in blue color were excluded during the experiment.*

The implication of the message design

Fact-checking outcomes may depend on intervention design. The first observation revealed that the borderline fact-checking messages might be perceived differently based on subtle changes in the wording; people reacted similarly to Lack of Evidence and Mostly False but not to the others (i.e., Mixed Evidence, Divided Evidence). Systems may postpone publishing decisions until sufficient data is gathered. However, such practice may also harm the system's credibility when untagged claims accumulate. Alternatively, fact-checking systems may provide interim results such as 'Emerging Issues,' 'High Complexity,' and 'Deliberately Unclear Claim.' When designing new tags, our methodology to expose fact-checked tags and compare the pre-stance and post-stance towards claims may be useful.

While not explored in the current study, some fact-checking outlets adopt comical meters, such as 'Pants on Fire' in PolitiFact. Such colorful tags may similarly incur other effects beyond what was intended to be conveyed. Conversely, in certain outlets, the middle verdicts like half-true are nonexistent, forcing online users to a decisive ruling. Hence, more user studies are needed to understand how different wordings and rating choices may affect the fact-checking outcome.

There may exist biases linked to definite messages such as Mostly True and Mostly False in contrast to borderline messages. Disconfirmation bias is a cognitive process that describes motivated reasoning (Taber & Lodge, 2006). If the information shown is opposite to one's initial view, this bias leads to considering the counter-evidence unfavorably and even refusing it (Edwards & Smith, 1996). With a doubting mind, multiple logic processes including as disconfirmation bias may be acting to interpret the information (Bersoff, 1999). The current work explores disbelief-activated disconfirmation bias, which is activated when an opposite fact-checking condition is provided to those who have an initially negative view towards the claim.

Our findings have direct implications for the future design of fact-checking services. We highlight fact-checking effects and shed light on what kind of intervention designs might be more effective. While the current fact-checking systems expose users to a mere fact, alternative models could be designed to build up recognition of the fact more gradually. This is because fact-checking results come under attack from critics who disagree with their verdicts (i.e., firm believers). By combining these observations, one may better understand the possible biases unintentionally triggered by fact-checking tags.

Findings

Hypotheses

- [$H1_a$ Risk Avoidance] Online users exposed to the fact-checking tag Divided Evidence are more likely to develop a negative stance toward an online claim than those exposed to Mixed Evidence. — REJECTED
- [$H1_b$ Uncertainty Avoidance] Online users exposed to the fact-checking tag Lack of Evidence are more likely to develop a negative stance toward a claim than those exposed to Mixed Evidence. — SUPPORTED
- [$H2$ Disbelief-activated Disconfirmation Bias] Online users who initially have a negative view of the claim are less likely to correct their views upon seeing a fact-checked result than those with a positive pre-stance. — SUPPORTED

Uncertainty avoidance

The average post-stance on rumors after seeing fact-checking conditions along with 95% confidence intervals (see Figure 2) shows that those survey respondents exposed to the Divided Evidence condition show no significant difference in post-stance than those exposed to the Mixed Evidence condition or no condition at all (i.e., None). In contrast, fact-checking messages that trigger risk-aversion behaviors (such as Divided Evidence) did not cause subjects to distrust an unproven claim more than presenting them with either a neutral result or no result at all. However, the Lack of Evidence condition displayed a far more negative post-stance than None or Mixed Evidence, close to the post-stance of the Mostly False condition. This finding suggests that acknowledging insufficient evidence of a rumor may have led subjects to lean towards distrusting the rumor (post-stance difference between None and Lack of Evidence is 0.21, P < .001, the difference is from post-hoc Tukey's Honest Significant Difference (HSD) Test at the group level ANOVA: F = 83.62, P < .001). In short, we observed uncertainty-aversion, but not risk-aversion.

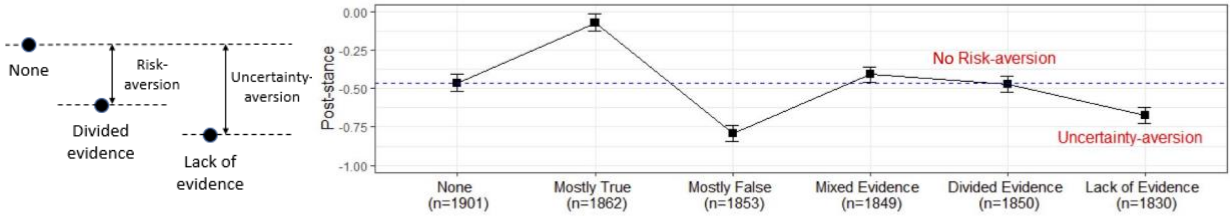


***Figure 2. Hypothetical effect of borderline fact-checking conditions (left diagram) and empirical results (right plot)**: 1) left: two psychological processes are depicted, where risk-aversion is represented as the value difference between post-stance of None and Divided Evidence conditions and uncertainty-aversion as the difference in post-stance of None and Lack of Evidence conditions, respectively; 2) right: post-stance on rumors after seeing a hypothetical fact-checking message (95% CI). The blue line shows the average post-stance of the None fact-checking condition, used as a baseline. The significance tests confirm not risk-aversion but uncertainty aversion. A five-point scale was used: (-2: definitely false, -1: false, 0: middle of the road, 1: true, 2: definitely true).*

We repeated the test over subgroups of users by their political leaning: Liberal (N = 3,758), Middle (N = 3,761), and Conservative (N = 2,008). We excluded participants who did not respond to this question. Based on the ANOVA followed by post-hoc Tukey's HSD tests, all three subgroups showed uncertainty-avoidance (Liberal: the post-stance difference between None and Lack of Evidence is 0.18, P = .08, Middle: 0.25, P < .01, Conservative: 0.28, P = 0.03). No subgroups showed risk avoidance. Interestingly, conservatives reported more significant negative changes when exposed to Lack of Evidence than other subgroups. This may imply that conservative young adults showed a higher degree of uncertainty-avoidance bias. This finding may be interpreted together with recent research that showed conservatives are more likely to respond to negative information online (Han et al., 2020).

Disbelief-activated disconfirmation bias

Next, we controlled for a single choice of the borderline condition by focusing on the subset of experiments when Mixed Evidence was presented along with all other conditions. We compared this result with None, the case when no fact-checking message was shown. We examined the changes in user stance depending on the four fact-checking conditions. Figure 3 shows the changes in participants' stance (the *y*-axis) towards rumors after seeing the fact-checking conditions (the *x*-axis). Each bar shows the 95% confidence interval, while each point denotes the mean. For the convenience of representing the stance change towards the given rumor claims, we assigned an integer variable to the five-point Likert scale and computed the scale difference before and after the intervention:

$$\textit{Stance change} = \textit{post-stance} - \textit{pre-stance}. \qquad (1)$$

Figure 3(a) shows that fact-checking systems affect users' stance to some extent; participants who were given the Mostly True condition had relatively positive stance changes towards the rumors than those given the Mostly False condition. The Mixed Evidence condition yielded a similar outcome as the No condition. We checked if the difference between the true and false conditions holds even after controlling for participants' pre-stance. For this, participants were grouped based on their pre-stance as positive (i.e., Definitely True or True) or negative (i.e., Definitely False or False).

The average stance change is below zero (i.e., negative direction) for all conditions for participants with positive pre-stance in Figure 3(b). In contrast, the stance change is above zero (i.e., positive direction) for those with negative pre-stance in Figure 3(c). This means that when young adults are exposed to the fact-checking message, they were less likely to take extreme views irrespective of the message.

What about when subjects were exposed to a fact-checking message opposite to their prior stance?

The figures indicate that stance change indeed happens in the direction guided by the fact-checking messages, denoted by the difference in the red dashed lines in Figure 3(b) and 3(c). Stance change direction follows the intervention message. Relatively, the negative pre-stance group was less likely to change their beliefs (i.e., the mean stance change difference between None and Mostly True is 0.59, P < .001, F = 157.8, P < .001) than the positive pre-stance group (i.e., stance change from None and Mostly False is 0.82, P < .001, F = 136.9, P < .001). This finding implies 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 is false) in combating online misinformation. Our finding confirms the existence of the disbelief-activated disconfirmation bias and emphasizes the importance of understanding human psychology in devising fact-checking schemes.

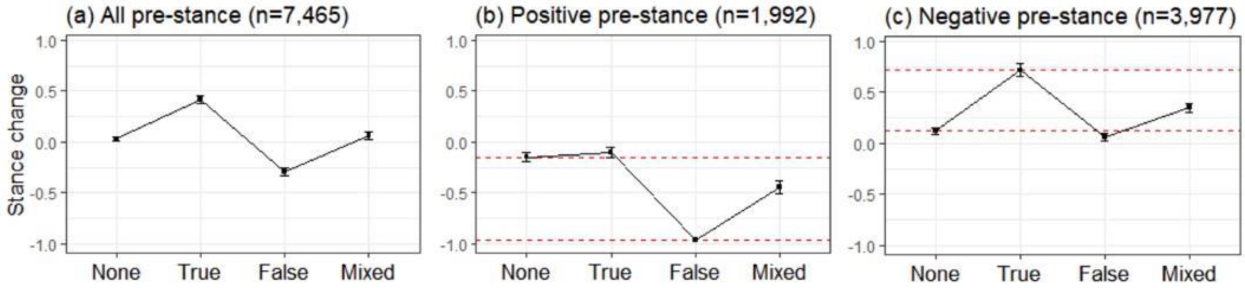


Figure 3. Stance change after seeing a fact-checking condition (95% CI). *The x-axis compares the experimental conditions: No condition, Mostly True, Mostly False, and Mixed Evidence. The y-axis indicates the magnitude of stance change in Equation (1). The two red lines represent the mean difference between two fact-checking conditions: None and the opposing tag to the pre-stance. The larger the gap between these lines, the larger the effect of the fact-checking message.*

Methods

We designed and conducted a set of randomized surveys to test the effects of fact-checking. We have borrowed the design frameworks from existing literature (Jun, Meng, & Johar, 2017; Babaei et al., 2019). The studies utilized the Amazon Mechanical Turk (MTurk) system to conduct experiments simulating online news consumption environments. We believe this method is acceptable in observing and imitating online human behaviors on a large scale, so we follow the norms introduced in those papers. The MTurk system is a well-known crowdsourcing platform frequently utilized to sample respondents and ask them to perform specific tasks for various research domains such as psychology, social science, and computer science (Strickland & Stoops, 2019).

Table 1. *Tested rumors drawn from the unconfirmed list of claims on Snopes.com.*

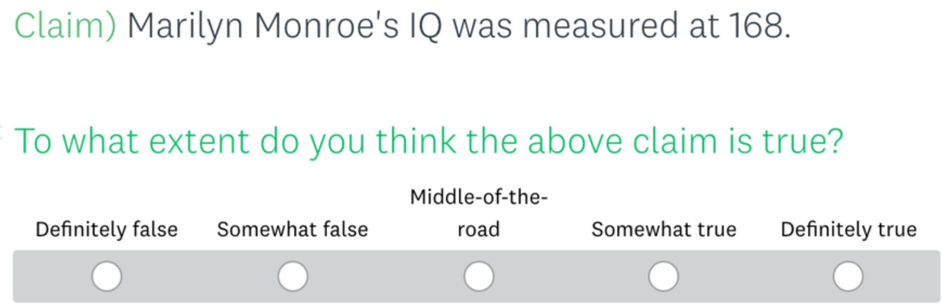
Category	Rumor
Food & Health	1. Vegetarians live longer than meat-eaters. 2. Drinking grape juice three times a day after being exposed to stomach flu will prevent you from sickness. 3. Coffee serves as an effective mosquito repellent and protection against infection by the Zika virus.
Politics	4. In 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 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.
Life & Entertainment	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.

Our surveys titled, "True or False? Rate an Unproven Claim," were intended to quantify online users' stance towards given rumors and their willingness to share. We selected ten rumors in Table 1 from Snopes.com using their 'unproven' claim category and choosing topics of interest to the general public. We selected rumors from diverse categories like food & health, politics, and life & entertainment to minimize topical biases. In terms of politics, rumor claims #4 and #5 are more favorable to conservatives, whereas #6 and #7 are to liberals. Given the limited set of rumors, our findings could be characterized as a suggestive rather than definitive study. Future studies can test and elaborate these findings over a broader set of rumors.

Our surveys showed participants a set of claims that are yet unproven. When presenting the ten claims, we do not exert any pre-assumptions towards these claims. 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 enrolled in the survey over a month (average age = 34 (18–36), 52.93% female). Each worker was allowed to participate only once and received a fixed amount of payment upon completing the study (i.e., the median time taken was 96 seconds with a payment of 7.5 USD per hour, which was guided by the minimum wage in the U.S. at the time of experiment in 2019). Personal information that we collected included the birth year, gender, ethnicity, and political orientation.

We recruited young adults who were 18 to 36 years old (at the time of the experiment) for the reasons below. First, many studies, including Baum et al. (2020), have shown that younger age groups tend to be more susceptible to online fake news. Second, younger adults are active in social media in posting and disseminating information (Ilakkuvan et al., 2019). Hence, examining how this cohort responds to fact-checks is of considerable importance.

Participants were asked to rate whether they think a given claim is true or false (hereafter *pre-stance*). Figure 4 illustrates an example of a randomly selected rumor and the question asking for pre-stance in the survey. The pre-stance was measured on a five-point 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 (Worcester & Downham, 1986). Next, we presented participants with one of the six fact-checking decisions chosen randomly from our system: 1) None for when 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, including three borderlines, we used graphical representations in Figure 1. Each experiment highlighted the target design in red and the rest in light grey.



Claim) Marilyn Monroe's IQ was measured at 168.

To what extent do you think the above claim is true?

Definitely false	Somewhat false	Middle-of-the-road	Somewhat true	Definitely true
○	○	○	○	○

Figure 4. An example rumor and a question asking for users' initial view on the claim.

After exposing a random fact-checking message, we then asked about the participants' stance toward the claim once again on a five-point scale (hereafter *post-stance*). The control group, who were shown no fact-checking message, were asked the identical set of questions on willingness to share and post-stance.

Our goal was to collect at least 150 responses for each of the six fact-checking conditions for ten rumors, which scales to 150 x 6 x 10 = 9,000 responses. The survey ran for a month-long period, which was ample time to reach this goal. Before analysis, any incomplete response or response with a duplicate IP address was removed. In the end, a total of 11,145 valid respondents participated in this experiment. Pre-stance leaned towards negative perception (negative = 53%, neutral = 20%, positive = 27%, P < .001, significantly different confirmed by proportions test) overall. The pre-stance distributions were not equal among the ten rumors (chi^2 = 1928.32, df = 36, P < .001, significantly different confirmed by Chi-square test), indicating that some claims were more likely to be believed than others.

Table 2 shows the participant counts' breakdown based on their pre-stance and post-stance on the ten claims. According to Snopes.com, counting the two extreme ends, 28.0% and 5.7% of the respondents said the rumors are Definitely False and Definitely True, respectively, although all stories were neither true nor false. Post-stance distribution is similar to that of pre-stance, yet fewer respondents then took extreme views.

***Table 2**. Summary of users' pre-stance and post-stance on 10 claims.*

Pre \ Post	Definitely False	False	Middle	True	Definitely True	Total
Definitely False	2,461	356	157	71	53	3,098
False	232	2,083	366	175	10	2,866
Middle	72	283	1,564	253	33	2,205
True	54	197	445	1,605	83	2,384
Definitely True	14	28	60	156	334	592
Total	2,833	2,947	2,592	2,260	513	11,145

Bibliography

Allington, D., Duffy, B., Wessely, S., Dhavan, N., & Rubin, J. (2020, June 9). Health-protective behavior, social media usage, and conspiracy belief during the COVID-19 public health emergency. *Psychological Medicine*, 1–7. https://doi.org/10.1017/S003329172000224X

Babaei, M., Chakraborty, A., Kulshrestha, J., Redmiles, E. M., Cha, M., & Gummadi, K. P. (2019). Analyzing biases in the perception of truth in news stories and their implications for fact-checking. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 139. https://doi.org/10.1145/3287560.3287581

Baum, M., Ognyanova, K., Chwe, H., Quintana, A., Perlis, R., Lazer, D., Druckman, J., Santillana, M., Lin, J., Volpe, J., Simonson, M., & Green, J. (2020). State of the nation: A 50-state COVID-19 survey: Report #14: Misinformation and vaccine acceptance. *Homeland Security Digital Library*. https://kateto.net/covid19/COVID19%20CONSORTIUM%20REPORT%2029%20ELECTION%20DEC%202020.pdf

Bersoff, D. M. (1999). Why good people sometimes do bad things: Motivated reasoning and unethical behavior. *Personality and Social Psychology Bulletin*, *25*(1), 28–39. https://doi.org/10.1177/0146167299025001003

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., Sandhu, M., Sang, R., Scholz-Bright, R., Welch, A.T., Wolff, A.G., Zhou, A. & Nyhan, B. (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, 1073-1095. https://doi.org/10.1007/S11109-019-09533-0

Edwards, K., & Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology*, *71*(1), 5–24. https://doi.org/10.1037/0022-3514.71.1.5

Epstein, L. G. (2004). A definition of uncertainty aversion. In I. Gilboa (Ed.), *Uncertainty in economic theory* (pp. 187–224). Routledge.

Hall, K. J., & Albarracín, D. (2020). The relation between media consumption and misinformation at the outset of the SARS-CoV-2 pandemic in the U.S. *Harvard Kennedy School (HKS) Misinformation Review*, *1*(3). https://doi.org/10.37016/mr-2020-012

Han, J., Cha, M., & Lee, W. (2020). Anger contributes to the spread of COVID-19 misinformation. *Harvard Kennedy School Misinformation Review*, *1*(3). https://doi.org/10.37016/mr-2020-39

Hassan, N., Arslan, F., Li, C., & Tremayne, M. (2017, August). Toward automated fact-checking: Detecting check-worthy factual claims by Claimbuster. *Proceedings of the 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* 1803–1812. https://doi.org/10.1145/3097983.3098131

Ilakkuvan, V., Johnson, A., Villanti, A. C., Evans, W. D., & Turner, M. (2019). Patterns of social media use and their relationship to health risks among young adults. *Journal of Adolescent Health*, *64*(2), 158–164. https://doi.org/10.1016/j.jadohealth.2018.06.025

Jun, Y., Meng, R., & Johar, G. V. (2017). Perceived social presence reduces fact-checking. *Proceedings of the National Academy of Sciences*, *114*(23), 5976–5981. https://doi.org/10.1073/pnas.1700175114

Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2013, December). Prominent features of rumor propagation in online social media. *Proceedings of the IEEE 13th International Conference on Data Mining*, 1103–1108. https://doi.org/10.1109/ICDM.2013.61

Stecula, D. A., Kuru, O., & Jamieson, K. H. (2020). How trust in experts and media use affect acceptance of common anti-vaccination claims. *Harvard Kennedy School (HKS) Misinformation Review*, *1*(1). https://doi.org/10.37016/mr-2020-007

Strickland, J. C., & Stoops, W. W. (2019). The use of crowdsourcing in addiction science research: Amazon Mechanical Turk. *Experimental and Clinical Psychopharmacology*, *27*(1), 1–18. https://doi.org/10.1037/pha0000235

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*(3), 755–769. https://doi.org/10.1111/j.1540-5907.2006.00214.x

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Worcester, R. M., & Downham, J. (1986). *Consumer market research handbook*. McGraw-Hill.

Zarocostas, J. (2020). How to fight an Infodemic. *World Report*, *395*(10225), 676. https://doi.org/10.1016/S0140-6736(20)30461-X

Funding

Meeyoung Cha was supported by the Institute for Basic Science (IBS-R029-C2) and the Basic Science Research Program through the National Research Foundation of South Korea (NRF-2017R1E1A1A01076400).

Competing interests

None.

Ethics

Our surveys stated that we are a group of researchers developing a fact-checking system to verify rumors. The participants' rights were protected throughout the research process, and this research was reviewed and approved by the Institutional Review Board (IRB) at the authors' institute. For every experiment, we obtained informed consent from the participants. At the end of the online survey, we debriefed the participants, informing them that the tested service is a mock system and that all fact-checking decisions had been randomly generated.

Copyright

Data Availability

All materials needed to replicate this study are available via the Harvard Dataverse:
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DHMLNR