# ONLINE FACT-CHECKER LABELS WORK—EVEN WITH THOSE WHO DISTRUST FACT-CHECKERS

## IN THIS BRIEF

- *Among the most widely used interventions against online misinformation are fact-checker warning labels. But do these warning labels actually work, especially with people who distrust fact-checkers? To find out, researchers Cameron Martel and David G. Rand ran a series of experiments.*
- *First, the researchers conducted a correlational study of 1,000 social media users to validate a measure of trust in fact-checkers.*
- *The researchers next conducted a total of 21 experiments with over 14,000 people. Participants were asked to evaluate both true and false news posts online. Those in the randomly selected treatment group saw warning labels on most of the false posts, while those in the control group saw no warning labels.*
- *The fact-checker warning labels worked. For the treatment group, warning labels reduced belief in labeled false information by nearly 28% and reduced misinformation-sharing by roughly 25% relative to a control group, which saw no warning labels.*
- *Among participants with less trust in fact-checkers, the reductions were smaller, but still significant. Warning labels for this group reduced belief in misinformation by nearly 13% and reduced misinformation-sharing by almost 17%.*

## OVERVIEW

Misinformation online presents a thorny dilemma. On the one hand, few responsible adults want to see the widespread dissemination of obviously false posts such as flat-earth theories, Covid-19 disinformation and immigrant fearmongering. On the other, social media platforms are protected by laws that essentially establish them as digital bulletin boards, freeing them of legal liability for the content their users post. What's more, managers of social media sites have been generally loath to moderate and remove all but the very worst misinformation. Social media sites are, among other things, businesses. For better or worse, misinformation drives traffic.

One promising intervention is the use of fact-checker warning labels. This occurs when professional fact-checkers find social media posts that are either false or misleading, and then mark these posts with their warnings (Figure 1).

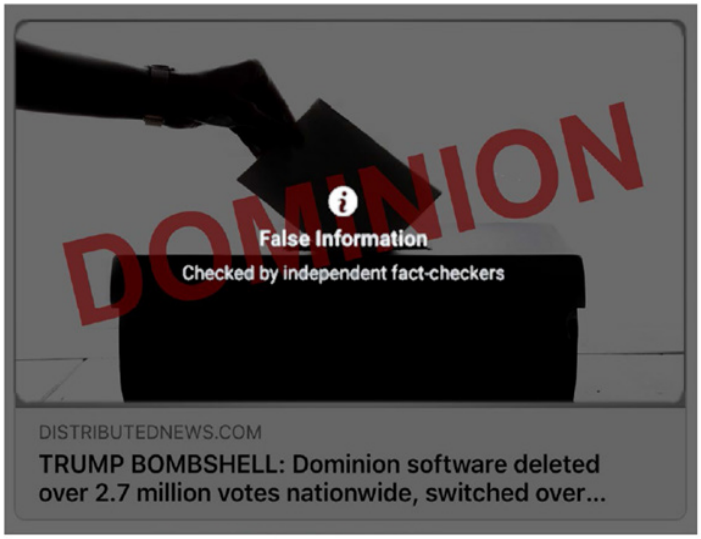


*Figure 1: Example of a false post on Facebook marked with a fact-checker's warning.*

Fact-checker warnings have been tried by some of the biggest social media sites, including Facebook, Instagram and Twitter/X. But do these warnings work, and do they actually discourage people from believing in and sharing misinformation? Equally important: Can these warnings work with people who inherently distrust fact-checkers?

Previous research (for example, Porter & Wood, 2022; Mena, 2020; and Pennycook et al., 2018) has suggested that fact-checker warnings do work, at least on average. But average effects may be of limited usefulness. Other studies (including Grinberg et al., 2019) have shown that exposure to online misinformation is mainly limited to comparatively small groups of people who follow low-quality domains.

Other studies show that in the United States, misinformation supply and sharing are concentrated among Republicans/conservatives, as other studies have shown (including Guess et al., 2019; and Guess et al., 2018). Still other research (Walker & Gottfried, 2019; Nyhan & Reifler, 2015) finds that Americans on the political right are substantially more distrustful of fact-checkers than those on the political left. This has led some researchers to suspect that fact-checker warning labels may not work—or could even backfire—with people who distrust fact-checkers.

To dig in beyond the averages, two MIT researchers—Cameron Martel and David G. Rand—in 2021 and 2022 conducted more than 20 online experiments involving over 14,000 social media users.

## THE EXPERIMENTS

The researchers first engaged 1,000 participants in a correlational study designed to assess the degree of trust in fact-checkers, which they abbreviate as TFC. There were three main findings. First, their eight-item TFC measure was reliable. It asks questions such as "How often do you think you can trust professional fact-checkers to check the news fairly?" This self-reporting measure also predicted whether participants would choose to see warning labels in a subsequent task. Second, the researchers confirmed that the more respondents leaned to the political right, the more distrustful of fact-checkers they were. And third, the researchers found that Republicans with greater news knowledge, analytic thinking and web-use skills had an even greater distrust of fact-checkers relative to Democrats.

Next, the researchers set out to determine whether distrust in fact-checkers undermines the efficacy of fact-checker warning labels. To do so, the researchers designed and conducted a total of 21 experiments involving 14,133 participants, all based in the United States. Ten of the experiments examined accuracy, while 11 examined sharing. The researchers conducted these experiments in four groups:

- **Experiment group 1:** These subjects were included in two experiments, one on accuracy, the other on sharing. Subjects were given 24 headlines to review, randomly selected from a list of 140 headlines that combined pro-Democrat, pro-Republican, and politically neutral messages. This experiment was conducted on Lucid, an online market-research tool.

- **Experiment group 2:** These subjects were also included in two experiments, one on accuracy, the other on sharing. And again, subjects were asked to review 24 headlines. Minor modifications were made from group 1. This experiment was also conducted on Lucid.

- **Experiment group 3:** This group participated in one experiment on sharing. Subjects were asked to review 36 headlines. This experiment was conducted on MTurk, a website that connects individuals and businesses.

- **Multiplatform experiments:** This group of 16 experiments—half on accuracy, half on sharing—was conducted across eight online recruitment platforms, including Connect, Forthright and Prolific. Subjects were asked to review 12 headlines selected from a pool of 108 headlines, a mix of pro-Democrat and pro-Republican messages.

## THE RESULTS

Overall, the experiments demonstrate that fact-checker warnings are generally effective at reducing both belief in and sharing of false headlines online. Among people who

distrust fact-checkers, the warnings were less effective, though the results were still significant.

In the first experiment, which tested a TFC measure with 1,000 people, the researchers validated that their measure is both highly reliable and predictive of who wants to view fact-checker warnings on false headlines. The researchers also replicated previous work demonstrating that Republicans are less trustful of fact-checkers (Figure 2).

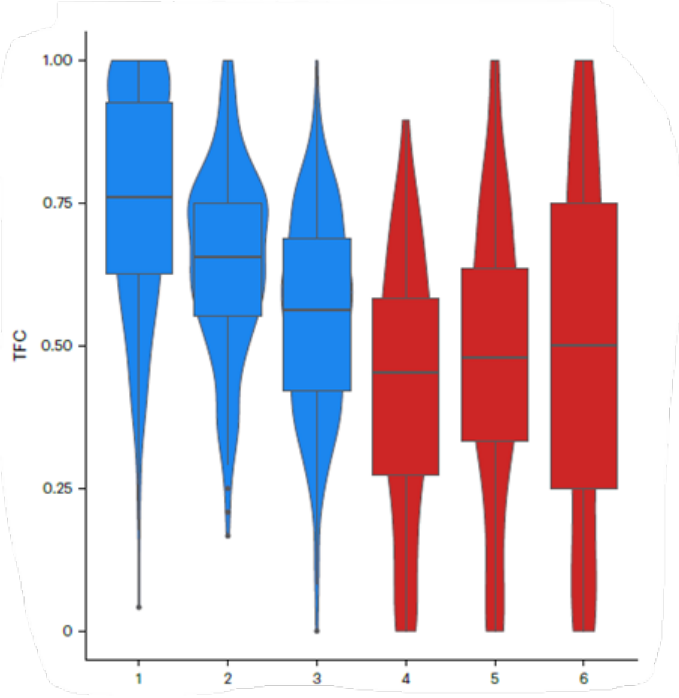


*Figure 2: The researchers found that Democrats (shown in blue) are more trusting of fact-checkers, while Republicans (red) are less trusting. On the X (horizontal) axis, 1 = strongly Democratic and 6 = strongly Republican. On the Y (vertical) axis, higher means more trust in fact-checkers (TFC), and lower means less. The red and blue "violin plots" show the distribution of experiment subjects across the two axes; wider means more people, and narrower means fewer.*

In the subsequent 21 experiments, subjects were asked to either rate the accuracy of false headlines or indicate how likely they would be to share the headlines. Participants in the treatment group were less believing of the labeled false headlines by a factor of 27.6% compared with the control group. Similarly, the treatment group was less willing than the control group to share false headlines by 24.6%.

Importantly, the warning effects persisted in subjects with low trust in fact-checkers. As stated above, the results were weaker than those for the overall group, but still significant: Belief in false headlines was reduced by 12.9% relative to the control group, and sharing of false headlines was reduced by 16.7%.

## CONCLUSIONS

Overall, the experiments illustrate what the researchers call a "discrepancy between self-reported attitudes and actual behavior." That is, people who said they distrust fact-checkers nonetheless reduced both their belief and sharing of labeled misinformation.

Why this gap between belief and behavior? The researchers offer five possible explanations:

- People can recognize that a specific headline is false, even as they maintain their general skepticism about fact-checkers.
- People may have been especially concerned about harming their reputations, outweighing their skepticism. Previous research (Altay et al., 2022) finds that sharing false information can hurt one's reputation. Sharing information previously labeled as false may send an even more negative reputational signal.
- Republicans' distrust of fact-checkers may have been mostly what the researchers call "expressive responding," rather than a true disbelief in the credibility of fact-checking. The response in this case was to either negative signals about fact-checking from Republican leaders or positive signals from Democrats.
- People with low TFC scores may both trust fact-checkers' skills and distrust how they select which headlines to examine.
- People who distrust fact-checkers may have other ways of recognizing when content is false, particularly when

they've been prompted by warnings. So even though these people distrust fact-checkers, being prompted by fact-checker warnings may lead them to consider a headline's other attributes.

The researchers say their work suggests warning labels can and should be put into practice for mitigating the effects of misinformation. And, as their experiments show that fact-checker warnings work even with people who distrust them, the researchers believe that concerns about warning labels backfiring are likely overstated.

## REPORT

Read the **full research report**.

## ABOUT THE RESEARCHERS

**Cameron Martel** is a doctoral candidate at the MIT Sloan School of Management. His research investigates why people believe and share misinformation, the forces that lead to online social network building, and interventions for content moderation.

**David G. Rand** is the Erwin H. Schell Professor and Professor of Management Science and Brain and Cognitive Sciences at MIT; an affiliate of the MIT Institute of Data, Systems and Society; Director of the Human Cooperation Laboratory and the Applied Cooperation Team; and Leader of the MIT Initiative on the Digital Economy's Misinformation & Fake News research group.

## REFERENCES


Altay, S., et al. (2022). **Why do so few people share fake news? It hurts their reputation**. New Media & Society, vol. 24, issue 6, pp. 1303-1324.

Grinberg, N., et al. (2019). **Fake news on Twitter during the 2016 U.S. presidential election**. Science, vol. 363, issue 6425, pp. 374-378.

Guess, A., et al. (2019). **Less than you think: Prevalence and predictors of fake news dissemination on Facebook**. Science Advances, vol. 5, no. 1.

Guess, A., et al. (2018). **Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 U.S. presidential campaign**. European Research Council; Brussels.

Mena, P. (2020). **Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook**. Policy & Internet, vol. 12, issue 2, pp. 165–183.

Nyhan, B. & Reifler, J. (2015). **Estimating fact-checking's effects: Evidence from a long-term experiment during campaign 2014**. American Press Institute; Arlington, Va.

Pennycook, G., et al. (2018). **Prior exposure increases perceived accuracy of fake news**. Journal of Experimental Psychology: General, vol. 147, no. 12, pp. 1865-1880.

Porter, E. & Wood, T.J. (2020). **Political misinformation and factual corrections on the Facebook news feed: Experimental evidence**. Journal of Politics, vol. 84, no. 3, pp. 1812–1817.

Walker, M. & Gottfried, J. (2019). **Republicans far more likely than Democrats to say fact-checkers tend to favor one side.** Pew Research Center; Washington, D.C.

MIT INITIATIVE ON THE DIGITAL ECONOMY

**MIT Initiative on the Digital Economy**
MIT Sloan School of Management
245 First St, Room E94-1521
Cambridge, MA 02142-1347
**ide.mit.edu**

**Our Mission:** The MIT Initiative on the Digital Economy (IDE) is shaping a brighter digital future. We conduct groundbreaking research on the promise--and peril--of new digital techonologies including generative artificial intelligence (GenAI), quantum computing, data analytics, and distributed marketplaces. We also investigate the rise of fake news and misinformation and the development of a digital culture. Through research and the convening of leaders from academia, industry, and government, the IDE provides critical, actionable insight for people, businesses, and government to understand and benefit from new technologies and how they're repidly changing the ways we live, work, and communicate.

**Contact Us:** David Verrill, Executive Director,
MIT Initiative on the Digital Economy
617-452-3216
**dverrill@mit.edu**

**Become a Sponsor:** The generous support of individuals, foundations, and corporations help to fuel cutting-edge research by MIT faculty and graduate students. It also enables new faculty hiring, curriculum development, events, and fellowships.

**Additional Contact:** Albert Scerbo, Associate Director,
MIT Initiative on the Digital Economy
267-980-2616
**ascerbo@mit.edu**

**View all our sponsors**

**Connect with us:**