Online misinformation warning labels work despite distrust of fact-checkers

Could online warning labels from fact-checkers be ineffective — or perhaps even backfire — for individuals who distrust fact-checkers? Across 21 experiments, we found that the answer is no: warning labels reduce belief in, and sharing of, posts labelled as false both on average and for participants who strongly distrust fact-checkers.

This is a summary of:
Martel, C. & Rand, D. G. Fact-checker warning labels are effective even for those who distrust fact-checkers. *Nat. Hum. Behav.* https://doi.org/10.1038/s41562-024-01973-x (2024).

Published online: 13 September 2024

The question

One of the most widely used interventions against online misinformation is attaching warning labels from professional fact-checkers to posts they have reviewed and found to be false. Emerging research suggests that such warning labels can reduce belief in, and sharing of, false posts on average[1]. However, average effects are of limited utility when assessing the effectiveness of anti-misinformation interventions, as exposure to online misinformation is extremely heterogeneous. In particular, in the US context, exposure to and sharing of false-news URLs is largely concentrated among Republican-favouring individuals from the political right[2]. Previous work suggests that these individuals also have a heightened distrust of fact-checkers[3]. This has raised concerns of warning labels potentially being ineffective or backfiring for this population. Thus, we set out to determine whether such a lack of trust in fact-checkers undermines the efficacy of fact-checker warning labels.

The observation

We first recruited a quota-matched US sample (n = 1,000) to validate a measure of trust in fact-checkers[4] and examine its correlates. Our 'trust in fact-checkers' measure predicted whether participants chose to see fact-checker warning labels on news evaluated as false by professional fact-checkers, thereby validating it. As expected, we also confirmed that participants with stronger Republican leanings were less trusting of fact-checkers. Next, we conducted 21 online experiments (total n = 14,133 participants) in which participants were asked to evaluate a series of true or false news posts, and were randomly assigned either to see no warning labels or to see warning labels on a high proportion of false posts. For each, participants either rated how accurate they thought the headline was — their perceived accuracy — or their intention to share it. We then meta-analysed the results of these experiments to assess the effects of warning labels on belief in, and sharing of, false posts across levels of trust in fact-checkers.

Our findings showed that warning labels are effective on average (Fig. 1). Warning labels reduced belief in (27.6% reduction) and sharing of (24.7% reduction) false posts. But did distrust in fact-checkers undermine the efficacy of warning labels? We found smaller effects of warning labels on accuracy perceptions for participants with less trust in fact-checkers, but that warning labels nonetheless substantially reduced belief in false posts (12.9% reduction) even for participants who were most distrusting of fact-checkers. We found similar results for sharing intentions. Although warning labels had nominally smaller effects for those less trusting of fact-checkers, warnings still reduced sharing of (16.7%) false posts for participants who strongly distrusted fact-checkers.

Future directions

Our results suggest that fact-checker warning labels are a broadly effective tool for combating misinformation. Although we observed somewhat stronger warning-label effects for those with greater reported trust in fact-checkers, we also found that warning label effects still largely persist across levels of fact-checker distrust. Our results illustrate an important instance of discrepancy between self-reported attitudes and actual behaviour — individuals who reported low trust in fact-checkers nonetheless reduced their belief in, and sharing of, false posts labelled by fact-checkers (Fig. 1). Our findings also assuage concerns about potential backfire effects of warning labels within populations of individuals more distrusting of fact-checkers.

Our studies have several limitations. First, all of our experiments showed participants politically balanced news feeds and showed participants warning labels on an equal number of pro-Democratic and pro-Republican party posts. Future work should investigate efficacy in environments with varying headline composition and labelling application. Second, our experiments used Facebook's current 'False Information' overlay as a warning label, and efficacy is likely to vary by warning label design. Last, future examinations should assess how warning labels function in social media environments with social feedback and user interaction. We also conducted all of our experiments with online convenience samples in the USA.

In future work, we plan to continue investigating potential heterogeneity in warning label effects and other misinformation intervention outcomes. Assessing how different interventions vary across individual differences (trust in fact-checkers and demographic factors), stimulus differences (high-plausibility versus low-plausibility posts), and environmental settings (high-quality versus low-quality news feeds) are important avenues for determining which strategies may be most effective for quelling misinformation across various contexts.

Cameron Martel & David G. Rand
Massachusetts Institute of Technology, Cambridge, MA, USA.

EXPERT OPINION

"The article advances our understanding of the effectiveness of fact-checks in the USA. Although some have worried that fact-checks could be ineffective or even backfire among those who do not trust fact-checkers, the article shows that these fears are not warranted. The scope of the article is impressive, but I cannot help wondering whether the results would replicate outside of experimental settings, when people who distrust fact-checkers can ignore fact-checks." **Sacha Altay, University of Zürich, Zürich, Switzerland.**

FIGURE

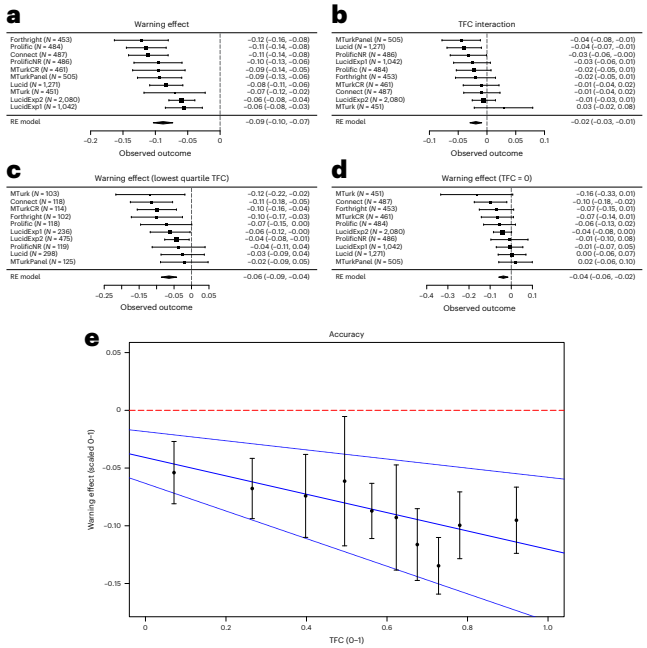

Fig. 1 | Warning labels reduce perceived accuracy of news labelled as false across individuals with varying levels of trust in fact-checkers. Participants who viewed a series of true or false news articles were randomly assigned to either see or not see warning labels. A random effects (RE) meta-analysis across multiple experiments (left columns) predicted perceived accuracy (PA) of labelled false posts relative to unlabelled false posts (in control) to estimate warning effects. **a**–**d**, Warning labels decreased PA of false content on average (**a**), were more effective for those with greater trust in fact-checkers (TFC) (**b**), decreased PA of false content for those in the lowest TFC quartile (**c**) and decreased PA of false content for those for whom TFC was minimal (TFC = 0) (**d**). **e**, Warning label efficacy on PA broadly persists across TFC levels (TFC of 0 to 1). © 2024, Martel, C. & Rand, D. G.

BEHIND THE PAPER

Decision-makers at several large technology platforms had expressed concern about whether warning labels from fact-checkers could perhaps be ineffective or even harmful — particularly if those who disliked fact-checkers were exposed to these warning labels. In previous work, we had assessed the effects of warning labels, but never examined the possibility of heterogeneous treatment effects by distrust in the source of these fact-checks. After our study validating a measure of fact-checker trust, we initially conducted two preregistered experiments examining warning label effects and moderation by trust in fact-checkers. We found average warning effects and no moderation by trust in fact-checkers. An associated research team then began conducting a multiplatform assessment comparing different online recruitment platform samples and, as part of this study, we were able to include a shortened version of our warning label experiment. This enabled us to perform a preregistered internal meta-analysis across 21 different experiments. **C.M. & D.G.R.**

REFERENCES

1. Martel, C. & Rand, D. G. Misinformation warning labels are widely effective: a review of warning effects and their moderating features. *Curr. Opin. Psychol.* **54**, 101710 (2023).
   A review that presents a summary of research examining the average effects of warning labels on false news.
2. Guess, A., Nagler, J. & Tucker, J. Less than you think: prevalence and predictors of fake news dissemination on Facebook. *Sci. Adv.* **5**, eaau4586 (2019).
   An analysis of Facebook data that shows that political conservatives are more likely than liberals or moderates to share articles from fake news domains in the USA.
3. Nyhan, B. & Reifler, J. *Estimating Fact-Checking's Effects* (American Press Institute, 2015).
   Survey data evidence that Republicans have less favourable views of fact-checking than Democrats in the USA.
4. Tsfati, Y. & Cappella, J. N. Do people watch what they do not trust?: Exploring the association between news media skepticism and exposure. *Commun. Res.* **30**, 504–529 (2003).
   Empirical work on media scepticism that developed a media scepticism scale, which we adapted for our 'trust in fact-checkers' measure.

FROM THE EDITOR

"Misinformation and how to fight it are topics of high practical and scientific relevance. Many online platforms have used warning labels from professional fact-checkers as an intervention against online misinformation. This article shows that these labels are effective, even when individuals distrust fact-checkers."
Editorial Team, *Nature Human Behaviour*.