REDUCING MISINFORMATION SHARING WITH ACCURACY PROMPTS

By Hause Lin, Haritz Garro, Nils Wernerfelt, Jesse Shore, Adam Hughes, Daniel Deisenroth, Nathaniel Barr, Adam Berinsky, Dean Eckles, Gordon Pennycook, David G. Rand

IN THIS BRIEF

- *Reducing the sharing of misinformation on social media is an important priority. One highly scalable option is content-neutral intervention, because it doesn't require specific and potentially biased fact-checks.*
- *This research provides the first large-scale evaluation of a content-neutral intervention on two social-media platforms: Facebook (Meta) and X (formerly known as Twitter).*
- *We subjected both platforms to randomized controlled trials. The main finding: Simple messages that remind people to think about accuracy—delivered to large numbers of people using digital advertisements—reduce misinformation sharing.*
- *On Meta/Facebook, just one hour after an accuracy prompt, we found a 2.6% reduction in the probability that misinformation would be shared by people who had shared misinformation in the prior week.*
- *On the X platform, we similarly found a 3.7% to 6.3% decrease in the probability of low-quality content sharing among active users who had shared misinformation before.*
- *The findings suggest that content-neutral interventions could complement existing content-specific interventions in reducing the spread of misinformation online.*

The proliferation of misinformation on social media is a source of great concern, and a large effort has been made to reduce it. Beyond that, sharing of fake news and other misinformation worsens the problem. The mainstay of current approaches to stem the spread is content-specific, such as the algorithmic demotion of flagged content. While these content-specific approaches have been found to be effective (Martel & Rand, 2023), they are not scalable enough to keep pace with the huge amount of content posted on social media. For example, in 2022 some 1.7 million pieces of content were posted on Meta/Facebook every minute (DOMO, 2023). At this scale, fact-checking becomes highly challenging, to say the least!

Further, content-specific interventions can be impossible to conduct on platforms that use stringent privacy protections such as end-to-end encryption. Another issue is that some Americans have expressed concern over the possibility of bias and over-enforcement of content-specific interventions (Jaimungal, 2020). Two cases about whether states or social media platforms should moderate content were recently **presented to the U.S. Supreme Court**.

These concerns have raised interest in interventions that are content-neutral and that get ahead of the problem by reducing the spread of misinformation before it goes viral. Indeed, there now exists a large body of survey-based experiments showing that content-neutral interventions can combat misinformation sharing (Epstein et al., 2021; Guess et al., 2020; Pennycook et al., 2020; Pennycook et al., 2021; Roozenbeek et al., 2022).

To be sure, counterarguments have been made against content-neutral interventions. Some researchers argue that

accuracy prompts can't work for posts involving "sacred values," those that are central to political identities (Pretus et al., 2023). Others point out that if a social-media user sincerely believes a demonstrably false claim, an accuracy prompt cannot be expected to have a meaningful effect on their sharing behavior.

THE TWO EXPERIMENTS

To assess the effect of content-neutral interventions on users' actual sharing behavior, we conducted two large-scale experiments using what are known as accuracy prompts. Essentially, these are online advertisements that simply remind social-media users to consider accuracy in general.

However, these earlier surveys involved subjects who knew they were part of an experiment. As a result, there was no evidence that content-neutral interventions could reduce misinformation sharing at scale "in the wild."

To help address this gap, our results are from of two large-scale randomized field studies. One was conducted in early 2023 on (Meta) Facebook by an industry team that included Meta employees. The other experiment was conducted in late 2021 and early 2022 on X—then still called Twitter—by a group of academic researchers. In contrast to earlier experiments, the subjects were not informed that they were part of an experiment. Both experiments used

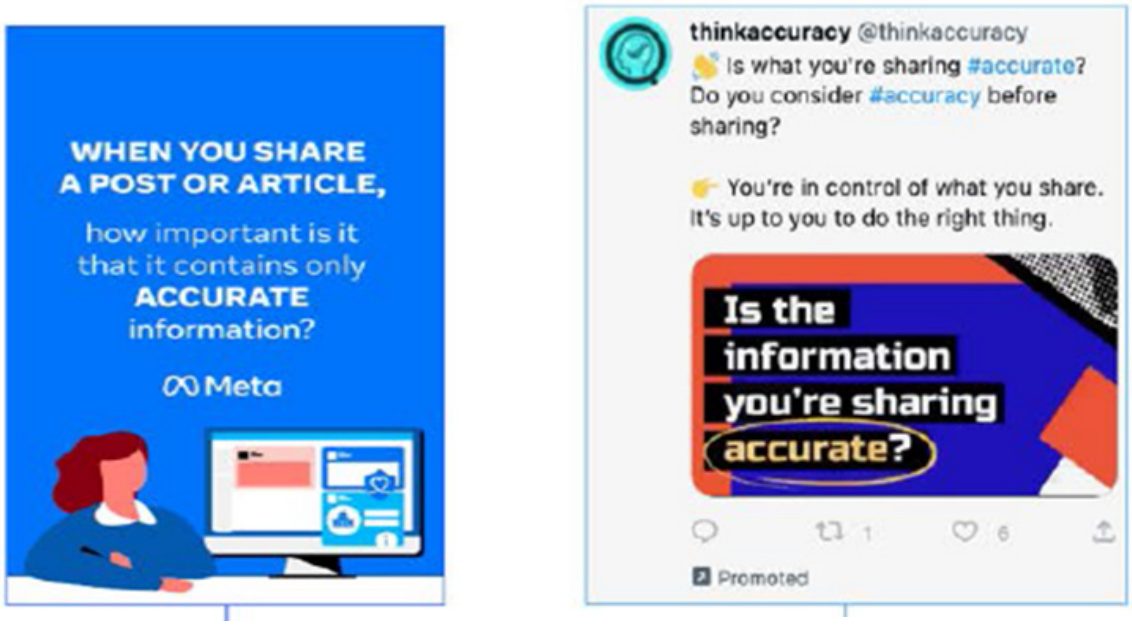


Figure 1. Accuracy prompt ads used in experiments on Facebook (left) and X (right).

Several survey experiments have shown that shifting users' attention back to accuracy can improve the quality of the content people intend to share online. For example, a recent meta-analysis of 20 survey experiments found that shifting attention to accuracy in various ways at the study outset reduced sharing intentions of false headlines by 10% (Pennycook & Rand, 2022). Another study, this one of over 34,000 people in 16 countries, found that accuracy prompts reduced sharing of false claims by 9.4% (Arechar et al., 2023).

advertisements to delivery accuracy prompts, and then they assessed the impact of these ads on the sharing behavior of users (*Figure 1*).

While online ads are a light-touch method—after all, users can simply scroll past them—they have been shown to affect both user behavior and attitudes (Gordon et al., 2022; Athey et al., 2023). Details of the experiments follow.

EXPERIMENT 1: ACCURACY PROMPTS ON META

The first experiment involved 33 million Meta/Facebook users, all selected by a random sample. Among them was a subset of users who, prior to the study, had repeatedly shared posts labeled as misinformation by fact-checkers.

We then assigned these users to either a treatment group or control group. Those in the treatment group were exposed to three accuracy prompts over three weeks; the ads were shown as replacements for regular ads. Those in the control group only viewed standard Meta/Facebook ads.

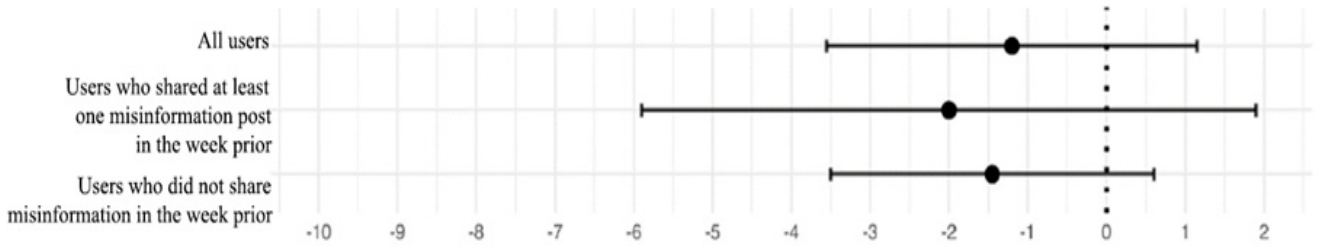
Users in the treatment group were further randomized to be prompted in three different ways: with static images; with a nine-second video; or with either a critical thinking image, a "poll ad," or a message stating that news accuracy is important.

We then measured the impact of these interventions during a 60-minute window after the first accuracy prompt was delivered (or, in the case of the control group, would have been delivered). Finally, we compared the sharing of posts containing misinformation by users in the treatment group with that of users in the control group.

As expected, the intervention reduced the number of users who shared misinformation by 1.8% relative to the control group. This was consistent across all three types of accuracy prompt treatments.

It's important to note that the intervention reduced sharing only among people who intended to share misinformation in the first place. Prior research shows this is a comparatively small fraction of all social-media users (Grinberg et al., 2019; Guess et al., 2019). Indeed, we found that most of our sample users (93.5%) had shared no misinformation in the week prior to the experiment. As expected, the effect of our intervention was greatest among those who had shared posts labeled as misinformation in the prior week (*Figure 2*).

Figure 2. Treatment effect on sharers of misinformation (absolute difference in basis points).



EXPERIMENT 2: ACCURACY PROMPTS ON X

This study, conducted by an academic research team, ran an ad campaign on the X platform that repeatedly showed users accuracy prompt ads. This also allowed us to evaluate whether the key insights from the initial Meta/Facebook experiment could be replicated and generalized.

As with the Meta/Facebook experiment, the X experiment randomly divided users into treatment and control groups. Users in the treatment group were exposed to roughly three daily video prompt ads over eight days. To keep the ad campaign interesting for users, the campaign involved a diverse set of 50 different video ads. Users in the control group were not exposed to the experimental ads.

We then conducted four experiments. Three of these experiments targeted highly active users who had recently shared links to low-quality news sites or questionable content. Two of these three experiments targeted users in the United States who had shared links related to "deep state" conspiracies. The third targeted users based in Canada who had shared hashtags linked to an anti-vaccination

protest. The fourth experiment targeted users who had not shared links to low-quality sites recently, but had previously done so.

We measured the amount of low-quality content shared by these users both before and during the ad campaign and then compared the actions of the treatment group with those of the control group. As expected, users in the treatment group shared 3.7% less low-quality content than did users in the control group.

However, we estimate that the effect was even greater. Due to privacy features of the X interface, our accuracy-prompt ads were not shown to all members of the treatment group, but only to 60% of them. We estimate that if all members of the treatment group had seen the ads, the average treatment effect would have resulted in a 6.3% reduction in misinformation-sharing.

FOUR KEY CONCLUSIONS

- Our experiments on Meta/Facebook and X show that accuracy prompts can reduce misinformation-sharing on social media platforms. This approach is promising because it's content-neutral; that is, information about specific posts is not needed to post the ads.

- Accuracy prompts can be used to help reduce the spread of misinformation. In this role, they can complement more traditional content-specific approaches such as fact-checking and the algorithmic identification of problematic content.

- The magnitudes of the effects we document—namely, 1.8% to 6.3% reductions in misinformation-sharing—are in line with our expectations, based on previous research. In addition, the similar responses we observed in both the Meta/Facebook and X experiments, despite their many differences in implementation, lend extra credence to our conclusions.

- For optimal effectiveness, these interventions must be frequent. Variation is important, too; a variety of ads keeps users engaged, a key to long-term success. We also find that these interventions are most effective when they're targeted at users who have shared misinformation in the past. By contrast, treating users who are not at risk is not cost-effective; it even has the potential for perverse effects.

REPORT

Read the **full working paper**

ABOUT THE AUTHORS

Hause Lin is a post-doctoral researcher at MIT Sloan School.

Haritz Garro is a Research Scientist at Meta Platforms.

Nils Wernerfelt is the Donald P. Jacobs Scholar and Assistant Professor of Marketing at Kellogg School of Management at Northwestern University.

Jesse Shore is a Research Scientist at Meta Platforms.

Adam Hughes leads the content policy research team at Meta.

Daniel Deisenroth is Director of Economics and Policy Research at Meta.

Nathaniel Barr is Professor of Creativity and Creative thinking at Sheridan College.

Adam Berinsky is the Mitsui Professor of Political Science at MIT and Director of the MIT Political Experiments Research Lab (PERL).

Dean Eckles is Associate Professor of Marketing at MIT Sloan and the lead for the MIT Initiative on the Digital Economy (IDE) New Data Analytics research group.

Gordon Pennycoock is Associate Professor of Psychology at Cornell University.

David G. Rand is the Erwin H. Schell Professor and Professor of Management Science and Brain and Cognitive Sciences at

MIT, and the lead for the IDE's Misinformation & Fake News research group.

ACKNOWLEDGEMENT

The academic research team thanks Antonio Arechar, Rocky Cole, Ziv Epstein, Beth Goldberg, Andrew Gully, Niko Lin, Mohsen Mosleh, and Michael Stagnaro for their helpful input, feedback, and comments.

REFERENCES

Arechar, A.A., et al. (2023). **Understanding and combatting misinformation across 16 countries on six continents**. Nature Human Behaviour, vol. 7, pp. 1502–1513.

Athey, S., et al. (2023). **Digital public health interventions at scale: The impact of social media advertising on beliefs and outcomes related to COVID vaccines**. Proceedings of the National Academy of Sciences (PNAS), vol. 120, no. 5.

DOMO (2023). **Data never sleeps 10.0**.

Epstein, Z., et al. (2021). **Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online**. Harvard Kennedy School Misinformation Review, vol. 2, no. 3, pp. 1–12.

Gordon, B. R., et al. (2022). **Close enough? A large-scale exploration of non-experimental approaches to advertising measurement**. Marketing Science, vol. 42, no. 4, pp. 768-793.

Grinberg, N., et al. (2019). **Fake news on Twitter during the 2016 U.S. presidential election**. Science, vol. 363, no. 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. (2020). **A digital media literacy intervention increases discernment between mainstream and false news in the United States and India**. Proceedings of the National Academy of Sciences (PNAS), vol. 117, no. 27, pp. 15536–15545.

Jaimungal, C. (2020). **America speaks: Do they think fact-checks of political speeches are helpful?** YouGov.

Martel, C., Rand, D.G. (2023). **Misinformation warning labels are widely effective: A review of warning effects and their moderating features**. Current Opinion in Psychology, vol. 54, pp. 1–10.

Pennycook, G., et al. (2020). **Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention**. Psychological Science, vol. 31, issue 7, pp. 770–780.

Pennycook, G., et al. (2021). **Shifting attention to accuracy can reduce misinformation online**. Nature, vol. 592, pp. 590–595.

Pennycook, G., Rand, D.G. (2022). **Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation**. Nature Communications, vol. 13, pp. 1–12.

Pretus, C., et al. (2023). **The role of political devotion in sharing partisan misinformation and resistance to fact-checking**. Journal of Experimental Psychology: General, vol. 152, no. 11, pp. 3116-3134.

Roozenbeek, J., et al. (2022). **Psychological inoculation improves resilience against misinformation on social media**. Science Advances, vol. 8, no. 34.

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: