Reducing misinformation sharing at scale using digital accuracy prompt ads

Hause Lin[1,2,10], Haritz Garro[3,10], Nils Wernerfelt[3,4], Jesse Shore[3], Adam Hughes[3], Daniel Deisenroth[3], Nathaniel Barr[5], Adam Berinsky[6], Dean Eckles[1,8], Gordon Pennycook[*2,7], & David G. Rand[*1,8,9]

[1]Sloan School of Management, Massachusetts Institute of Technology, [2]Department of Psychology, Cornell University, [3]Meta Platforms, [4]Kellogg School of Management, Northwestern University, [5]School of Humanities and Creativity, Sheridan College, [6]Department of Political Science, Massachusetts Institute of Technology, [7]Hill/Levene Schools of Business, University of Regina, [8]Institute for Data, Systems, and Society, Massachusetts Institute of Technology, [9]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

[10]These authors contributed equally: H. Lin, H. Garro.
*Corresponding authors: gordon.pennycook@cornell.edu, drand@mit.edu

Interventions to reduce misinformation sharing have been a major focus in recent years. Developing "content-neutral" interventions that do not require specific fact-checks or warnings related to individual false claims is particularly important in developing scalable solutions. Here, we provide the first evaluations of a content-neutral intervention to reduce misinformation sharing conducted at scale in the field. Specifically, across two on-platform randomized controlled trials, one on Meta's Facebook (N=33,043,471) and the other on Twitter (N=75,763), we find that simple messages reminding people to think about accuracy—delivered to large numbers of users using digital advertisements—reduce misinformation sharing, with effect sizes on par with what is typically observed in digital advertising experiments. On Facebook, in the hour after receiving an accuracy prompt ad, we found a 2.6% reduction in the probability of being a misinformation sharer among users who had shared misinformation the week prior to the experiment. On Twitter, over more than a week of receiving 3 accuracy prompt ads per day, we similarly found a 3.7% to 6.3% decrease in the probability of sharing low-quality content among active users who shared misinformation pre-treatment. These findings suggest that content-neutral interventions that prompt users to consider accuracy have the potential to complement existing content-specific interventions in reducing the spread of misinformation online.

This working paper has not yet been peer reviewed

The spread of misinformation is a source of great concern among academics, policymakers, and the general public[1]. Particular attention has been paid to social media's role in the spread of false and misleading content[2]. Accordingly, academics and technology companies have invested a great deal of effort in exploring approaches to reduce the spread of misinformation online. Labeling and algorithmic demotion of content flagged by machine learning classifiers, professional fact-checkers, or crowdsourced (i.e., layperson) evaluations[3] form the mainstay of current approaches to curb misinformation. These content-specific approaches have been shown to be largely effective at curtailing the influence of misinformation once it has been identified[4] (see, for example, an analysis of sharing before versus after posts were identified to Facebook as misinformation by 3rd party fact-checkers in SI section S1.3).

Critically, however, content-specific interventions alone cannot keep pace with the vast quantity of content posted on social media. For example, in 2022, 1.7 million pieces of content were posted on Facebook every minute[5]—fact-checking at this scale and pace would be challenging for any organization, particularly during crisis events. Furthermore, content-specific interventions are impossible on platforms with privacy protections such as end-to-encryption. Alternatively, some have expressed concern about the possibility of bias and over-enforcement of content-specific interventions, typically applied by platforms in a top-down fashion[6]. Thus, it is important to complement these traditional approaches with *content-neutral* interventions that get ahead of the problem by reducing the spread of misinformation before it "goes viral."

A large body of survey-based experiments suggests that content-neutral interventions may have promise for combating misinformation sharing[7-11]. However, virtually all this evidence relies on hypothetical sharing intentions measured while participants know they are in an experiment. Thus, despite an explosion of research in recent years, there is still virtually no evidence assessing whether content-neutral interventions can actually reduce the sharing of misinformation at scale "in the wild."

Here, we help to fill this gap. We focus in particular on *accuracy prompts*, a recently proposed class of intervention that is particularly scalable because it involves simply reminding users to consider accuracy. Even though the large majority of social media users around the world explicitly prefer to prioritize accuracy over other motives for sharing[10,12], the social media context can distract people from accuracy and focus their attention on other factors[9,10,13]. As a result, several survey experiments have shown that shifting attention back to accuracy can improve the quality of content people intend to share online[7,9,10,12-17]. In particular, having participants explicitly rate each headline's accuracy immediate before deciding whether to share it reduced sharing of false news by 28.4%-51.2%[10,13], and more generally prompting participants to consider accuracy at the outset of the study reduced subsequent sharing of false news by 10% in a meta-analysis of over 26,000 Americans[16] and by 9.4% in a study of over 34,000 people across 16 countries[12].

But can accuracy prompts reduce actual misinformation sharing on-platform? Some have argued that accuracy prompts are ineffective for Republicans[18] (but see[16]). Others have argued that accuracy prompts will not work for posts involving "sacred values" (i.e., those central to political identities) and that these posts are central to misinformation spreading[19]. Relatedly, it has been robustly demonstrated that accuracy prompts only reduce sharing insomuch as people believe the content is inaccurate[7,9,10,12,13,16]. Thus, if most misinformation shared online is more plausible than the blatantly false claims used in most survey experiments—or is shared by partisans who

sincerely believe the false claims—then we would not expect accuracy prompts to have a meaningful effect in the field.

Thus, despite a wealth of survey experiments, it remains unclear whether accuracy prompts are indeed a useful tool for platforms seeking to reduce the spread of misinformation[1]. To help address this gap, we present the results of two similar large-scale randomized field studies conducted by separate teams—one industry and one academic—on Facebook and Twitter, respectively, providing the first evaluations of accuracy prompts deployed on-platform at scale.

Both experiments use advertisements to deliver accuracy prompts and then assess the effect of these ads on users' actual sharing behavior. Although social media advertisements are a very light-touch method for delivering interventions—for example, users may scroll past without noticing the ads or may choose to ignore them purposely—they have nonetheless been shown to affect users' behavior and attitudes. For example, a meta-analysis of more than 600 advertising experiments using large advertisers on Facebook finds a median increase of 5% in purchase behavior and related outcomes[20]; a meta-analysis of over 800 public health advertising experiments on Facebook and Instagram found the average campaign increased the prevalence of positive opinions about COVID-19 vaccines by roughly 1%[21]; a long-term political advertising experiment during the 2020 U.S. Presidential Election caused a 0.3-0.4pp change in voter turnout[22]; and 90s video ads on YouTube describing rhetorical manipulation techniques increased technique identification in a follow-up survey by 5%[11]. Effects of this magnitude can have meaningful real-world consequences when deployed at scale (as suggested, for example, by the vast investment in digital ads by businesses and political campaigns).

Here, we ask whether accuracy prompt advertisements—delivered in a content-neutral manner—can similarly reduce users' subsequent sharing of misinformation. If so, we would demonstrate the utility of accuracy prompts for helping to address the misinformation challenge, provide evidence for inadvertent sharing of inaccurate content online, and offer guidance on how such content-neutral approaches may be most effectively delivered and applied.

Accuracy prompts on Facebook

Our first study (Fig. 1), an experiment conducted by a research team at Meta, involved 33 million Facebook users. The experiment primarily included non-targeted users reached through the ads auction, augmented with a smaller group of users targeted because they had repeatedly shared links classified as misinformation prior to the study (with ads also delivered via the ads auction). Users were randomly assigned to either a treatment group, whereby an average of 3.2 ads over the course of 3 weeks were replaced with accuracy prompts, or a control group, whereby those ad spots were filled with standard Facebook ad content.

[1] The only existing field evidence comes from a comparatively small-scale experiment in which Twitter users sent private accuracy prompt messages to their followers[10], which does not reflect how such prompts would be used by platforms.

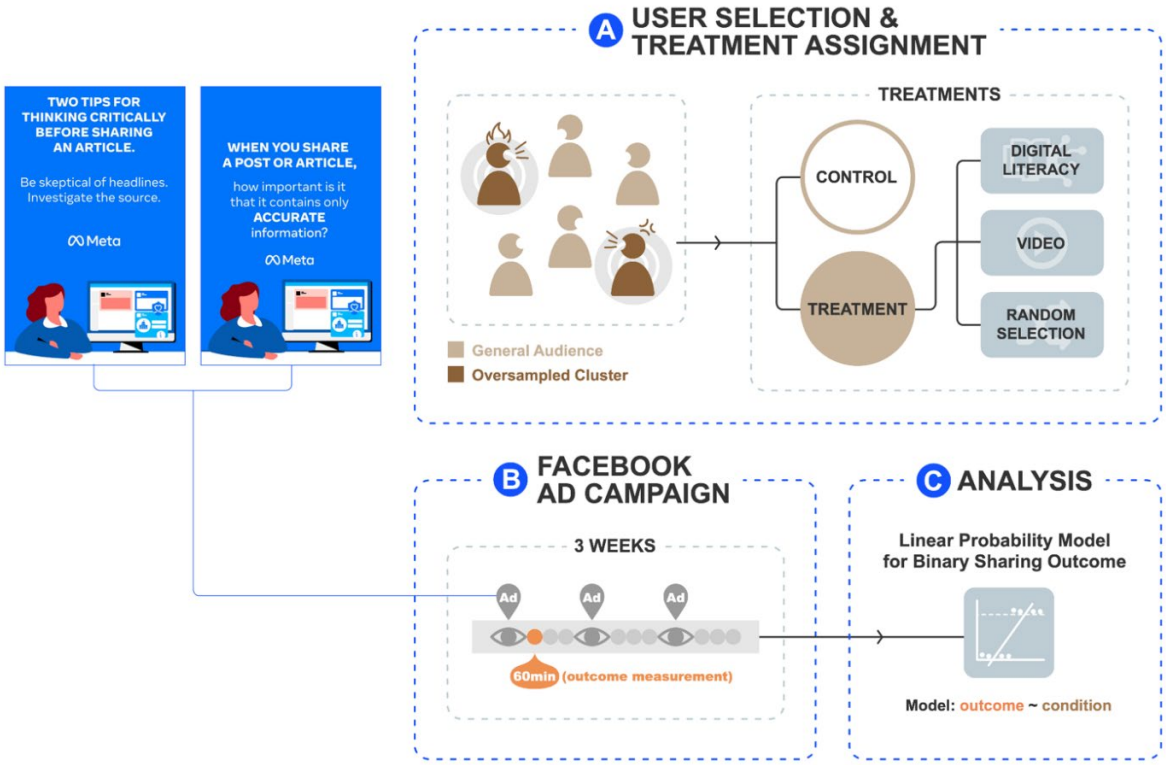

Fig. 1. Facebook study experimental design. *The Facebook study consists of two audiences where an oversampled cluster targeted users who had recently shared misinformation. Users were randomly assigned to the control group or one of three treatment groups. The intervention's impact was measured during the 60-minute window after the first accuracy prompt ad was delivered (or would have been delivered if the user was in the control group).*

Users in the treatment group were further randomized to be prompted to think about accuracy in three different ways: 1) static images providing "tips for thinking critically before sharing," 2) a 9-second video stating that some stories use emotional language and encouraging users to "check for accuracy," or 3) an ad randomly selected from either the critical thinking tips image, a "poll ad" asking users how important accuracy is when sharing, or a message that said that 88% of Americans believe it is important that the news they read online is accurate. Prior survey experiments have demonstrated that these different interventions are all similarly effective at shifting attention to accuracy[7].

We then compare the sharing of posts containing likely misinformation by users in the control versus the treatment. We classify a post as containing likely misinformation if it was flagged by third-party fact-checkers, a representative group of Facebook users[23], or Facebook's classifier that is trained to detect posts that are likely to be misinformation based on various signals[24]. Although this approach provides only an *estimate* of whether a given post contains misinformation, by using post-level evaluations, we achieve much greater precision than the domain-level evaluations typically used in prior work[10,25-27]. Furthermore, our approach achieves substantially greater coverage by being able to evaluate *all* posts rather than only posts containing links to news sites. Hereafter, we may refer to content flagged via these channels as

"misinformation" for convenience; for full details of the experimental design, see the Methods section.

Using this measure, we find that 5.7% of the non-targeted audience in our experiment shared at least one likely misinformation post in the week prior to the experiment, as did 74.6% of the users targeted because of past misinformation sharing. Overall, then, 6.4% of users in our experiment shared content estimated to be likely misinformation in the week prior to the experiment. Of course, the intervention could only possibly reduce misinformation sharing among users who had the opportunity and would have been inclined to share misinformation in the first place—and, as this analysis shows, despite our targeting a comparatively small fraction of users in our experiment share content estimated to be likely misinformation at baseline (in line with estimates from past work examining smaller samples of users with coarser measurements of information quality[28,29]).

Furthermore, the majority of users who did shared any likely misinformation shared only a single misinformation post (i.e., 60.6% when examining baseline data from the week prior to the study). Thus, our main specifications use a linear probability model to predict whether or not a given user shared any misinformation posts during the 60-minute post-treatment period, using a treatment dummy (see SI section S1.4 for models predicting the number of misinformation posts, which produces qualitatively equivalent results). For readability, we express all linear regression coefficients in units of basis points. For all null results, the larger magnitude endpoint of the reported 95% confidence interval corresponds to a 97.5% equivalence bound from a two one-sided test procedure—that is, for a null result with 95% CI of [a, b], the 97.5% equivalence bound is [-c, c] with c = max(|a|, |b|).

What, then, was the causal effect of receiving an accuracy prompt on the Facebook users' subsequent misinformation sharing? We begin by assessing the intervention's impact during the Facebook session in which the accuracy prompt was delivered by examining misinformation sharing in the hour after receiving the first accuracy prompt ad versus control ad. Across all users, we find that the treatment significantly reduced the number of users who shared misinformation posts (1.8% reduction relative to control; b = -0.38, 95% CI [-0.69, -0.07], p = .018; see Fig. 1A). A s expected, a significant interaction between treatment and pre-experiment sharing (b = -5.68, 95% CI [-9.60, -1.77], p = .004) shows that the treatment had a significantly larger effect among users who shared at least one misinformation post in the week before the experiment (2.6% reduction relative to control; b = -5.70, 95% CI [-9.62, -1.78], p = .004), compared to the users who did not (b = -0.02, 95% CI [-0.22, 0.18], p = .819). Thus, we focus our subsequent analyses on the 6.4% of users in the study who had shared misinformation in the week prior to the experiment (for parallel analyses including all users, see SI section S1.4).

Similarly, many users were likely not exposed to any misinformation during the hour after the prompt ad, such that the treatment could not have reduced misinformation reposting among these users—which in turn deflates our estimate of treatment's effect. While we do not have event-level misinformation exposure data for our experiment, a survey in which 42% of Facebook users claimed that they saw false content every time or almost every time they logged onto the platform[30] can serve as an illustrative upper bound. Such self-report measures are well known to dramatically overestimate exposure[31,32], and thus it is unlikely that 42% of users were actually exposed to misinformation—especially during the one hour period on which we focus our analyses. This illustrative upper bound would imply that the treatment reduced the fraction of

misinformation sharers by substantially more than 4.3% among all users who were exposed to misinformation during the experiment, and substantially more than 6.2% among users who shared misinformation in the pre-treatment period and were exposed to misinformation during the experiment. For example, if 20% of users were exposed, that would imply 9% and 13% reductions in misinformation sharing among exposed users (all users and those who shared misinformation pre-treatment, respectively), and if 10% of users were exposed, that would imply reductions of 18% and 26%.

Turning to potential moderators, we find no significant differences in effect size across the three different kinds of accuracy prompt ad treatments (digital literacy tip image = 1.0% decrease, 9s video = 3.4% decrease, random selection = 3.4% decrease; Wald test for joint significance, p = .234). This finding is consistent with the results of a survey experiment that compared numerous different accuracy prompts and suggests that the prompts are all working through a similar mechanism of shifting attention to accuracy[7,12]. More direct evidence for this mechanism comes from a post-experimental survey (administered using Meta's "brand lift" machinery), which suggested that treatment users reported thinking more about the accuracy of the posts they read on Facebook compared to control users (see SI section S1.5).

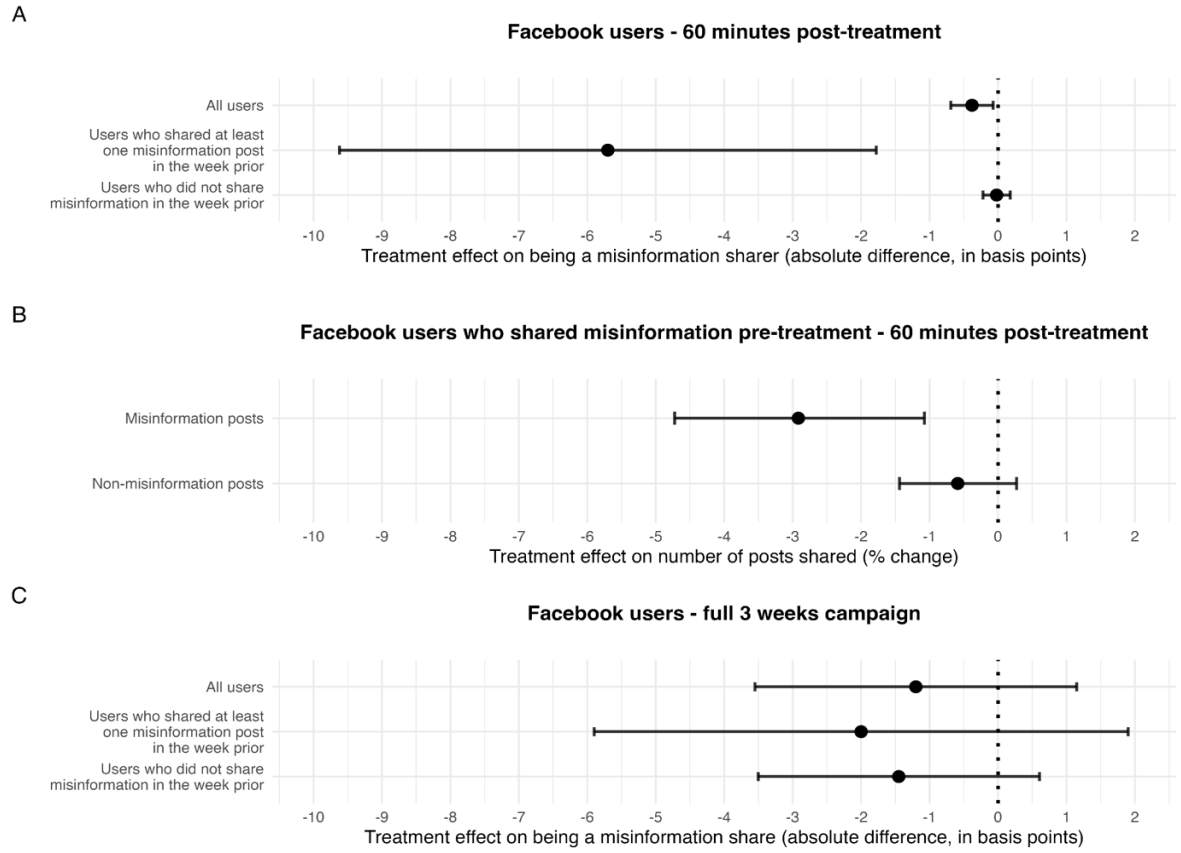

Fig. 2. A single accuracy prompt ad reduced misinformation sharing within the browsing session on Facebook*. **(A)** Estimates (in units of basis points) from linear probability models predicting whether or not a given user shared any misinformation posts during the 60-minute post-treatment period. **(B)** Estimates (in units of percent change) from quasi-Poisson models predicting the number of misinformation posts shared during the 60-minute post-treatment*

period, specifically for only users who shared misinformation at least one misinformation post in the week prior to the experiment. ***(C)*** *Estimates (in units of basis points) from linear probability models predicting whether or not a given user shared any misinformation posts over the full 3 weeks of the experiment. 95% confidence intervals are shown.*

Finally, we find no significant interactions between treatment and user age (>65 dummy; b = -8.0, 95% CI [-22.0, 5.7], p = .264), gender (female dummy; b = 2.00, 95% CI [-5.80, 9.80], p = .688), or education (college degree dummy; b = -2.00, 95% CI [-9.80, 5.80], p = .660). Importantly, we also found no significant effect of the treatment on the number of non-misinformation posts shared across 60 minutes (quasi-Poisson regression, b = -0.0059, 95% CI [-.0145, .0027], p = .180; compared to a significant effect on the number of misinformation posts shared, b = -0.0296, 95% CI [-.0484, -.0108], p = .002, which is significantly more negative than the effect on non-misinformation posts, p = .023; Fig. 2B). Thus, the treatment effect is likely specific to misinformation, rather than reducing sharing more generally[33]. This result is important from both a theoretical perspective, as the intervention was predicted to impact misinformation specifically, as well as a practical perspective, as platforms could be reluctant to implement anti-misinformation policies that reduce engagement with non-misinformation content.

We now turn to misinformation sharing over the full 3 weeks of the experiment. Given that digital ads in general have modest effect sizes[20-22] and sharing over 3 weeks is a high-variance behavior, and that accuracy prompts in particular rely on attention, which is easily redirected[2,34,35], we unsurprisingly did not find any significant treatment effect on the probability of being a misinformation sharer between the treatment and control when including all posts shared over the 3 experimental weeks (all users: b = -1.20, 95% CI [-3.60, 1.20], p = .313; users who shared misinformation pre-treatment: b = -2.00, 95% CI [-5.90, 1.90], p = .212; Fig. 2C).

Together, these results indicate that accuracy prompts *can* reduce misinformation sharing following exposure to the ad but that just a single or few prompts per week is not enough to produce a detectable overall effect. These observations suggest that more consistent redirection of attention to accuracy is required for a sustained reduction in misinformation sharing.

Accuracy prompts on Twitter

This proposition is evaluated in our second study, conducted by an academic research team, which involved conducting an ad *campaign* on Twitter that repeatedly showed users accuracy prompt ads. Our second experiment also allows us to evaluate the replicability and generalizability of the key insights from the first experiment by testing a very similar intervention on a different social media platform and using different implementation and operationalization details.

In our second study (Fig. 3), users were assigned to treatment or control via block randomization, and users in the treatment were used as the custom audience for an ad campaign that showed them an average of 2.91 accuracy prompt ads per day for at least 8 days. To avoid users beginning to ignore the prompts over repeated exposures, the ad campaign used a diverse set of

50 different accuracy prompt video creatives, all of which looked different but mentioned accuracy in some way (see SI section S2.1.4).

Using the exact same ad campaign setup and creatives, we conducted four separate experiments on different populations of Twitter users (Table 2). Three experiments targeted highly active Twitter users who had recently shared links to low-quality news sites or potentially problematic or questionable content prior to the study. Two of these three populations were largely U.S. users, selected based on having shared links to low-quality domains or tweets about deep-state conspiracies; the third population consisted of largely Canadian users, selected based on sharing hashtags linked to an anti-vaccination protest in Ottawa. A fourth experiment targeted users who had *not* recently shared links to low-quality new sites but who had done so further in the past.

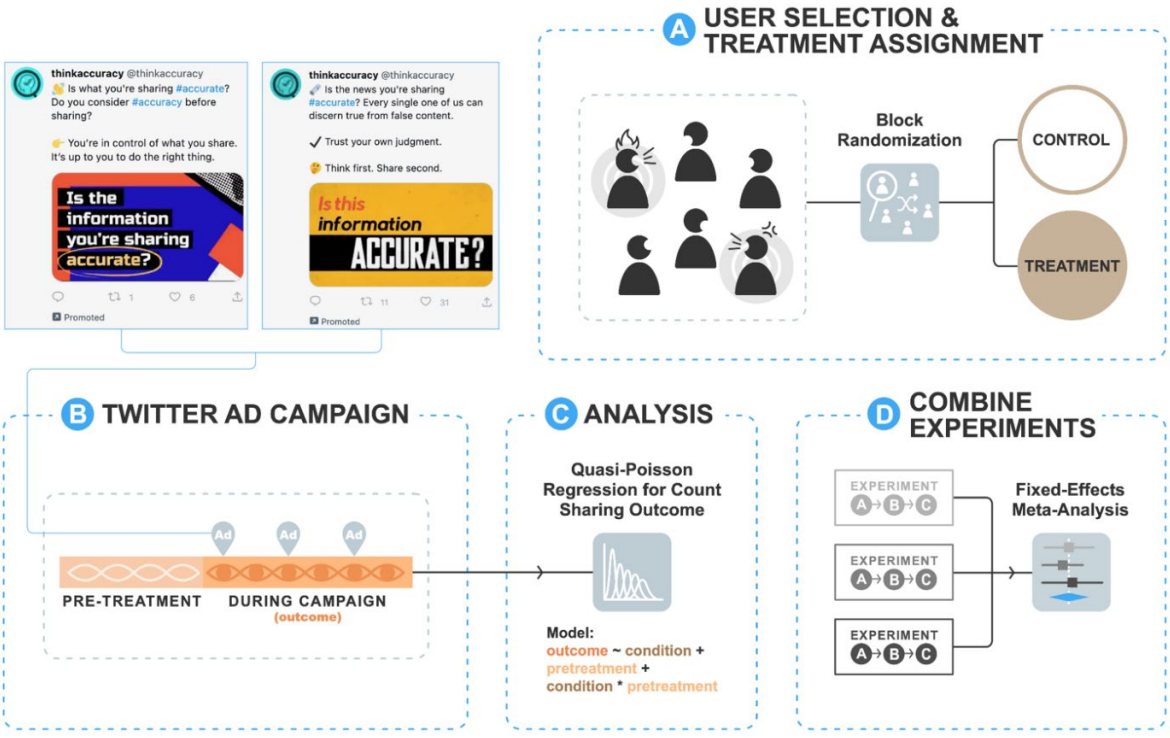

Fig. 3. Twitter study experimental design. *The Twitter study targeted highly active users who recently shared low-quality content. The amount of low-quality content shared by users was measured before and during the ad campaign. Results from three separate experiments were combined using fixed-effects meta-analysis.*

We then compared the amount of misinformation shared (i.e., retweeted without comment) among users in the treatment versus the control. To classify retweets as misinformation, we use the standard approach in the academic literature of using domain-level quality ratings[27]; our main analyses classify all retweets containing links to sites with quality scores at or below 0.70 on a 0-1 scale as "low quality" (in SI sections S2.2.2 and S2.2.5, we show that our results are robust to using alternative quality thresholds and also show results from an alternative more graded approach to scoring domain quality). While domain-level quality ratings are necessarily coarse, it is the only tractable approach given the large number of tweets. An exception, however, arises in the experiment targeting users who shared hashtags linked to an anti-vaccination protest, which *does* offer a tractable approach: counting the number of relevant hashtags shared by users. Thus,

for greater precision, our main pre-registered analyses for this experiment use hashtag counts instead of low-quality domain counts for that specific experiment (although results are similar when using low-quality domain counts for all experiments; see SI section S2.2.3).

We use quasi-Poisson regression to predict the number of low-quality posts (or hashtags) shared by each user on each day of the experiment, with condition as the independent variable and including block and date fixed effects and controlling for the users' level of pre-experiment misinformation sharing, as well as clustering standard errors on randomization block.

We begin by examining the three experiments that targeted highly active misinformation sharers because (as discussed and demonstrated above) we would only expect to find treatment effects on users who would have shared misinformation in the absence of the treatment. Our main analyses calculate overall effect sizes using fixed-effects meta-analysis across experiments. As expected (Fig. 4A), users in the treatment shared less low-quality content than users in the control (3.7% reduction relative to control; b = -0.038, 95% CI [-0.062, -0.013], p = .002).

Fig. 4. The accuracy prompt ad campaign reduced sharing of low-quality content among active misinformation sharers on Twitter. *The overall effect is the intent to treat (ITT) effect and the effect among users who saw at least 1 ad is the average treatment effect on the treated (ATT).* ***(A)*** *Estimates (black circles) are obtained by performing fixed-effects meta-analyses of three experiments recruiting active misinformation sharers. The experiment-level estimates (blue hollow circles) are coefficients (in units of percent change) from quasi-Poisson models predicting the number of low-quality (top) or non-low-quality (bottom) posts shared in each experiment.* ***(B)***

Estimates from quasi-Poisson models predicting the number of low-quality content posts shared in the experiment recruiting inactive misinformation sharers. 95% confidence intervals are shown.

Importantly, due to the nature of Twitter ad delivery, only 60% of the treatment users were shown any accuracy prompt ads at all (due to privacy features of the Twitter ads interface, we do not know which specific treated users did not see any ads, only the overall percentage). Thus, in addition to evaluating the overall *intent to treat* (ITT) effect reported above, we also estimate the causal effect of the treatment among those who actually received it using an instrumental variables approach in which we use the initial assignment to treatment as an instrument for actual treatment receipt. Using this approach, we estimate an *average treatment effect on the treated* (ATT; i.e., the effect of the treatment among users who would actually receive the treatment), which quantifies the true effect of being treated—which yields an estimate of b = -0.065, 95% CI [-0.107, -0.023], p = .002 (i.e., 6.3% reduction in misinformation sharing in treatment relative to control; Fig. 4A).

We next turn to examining moderators of the treatment effect. There was no strong evidence that the treatment effect varied based on users' pre-treatment level of misinformation sharing (interaction between treatment dummy and sqrt[number of misinformation posts in week pre-treatment]: b = 0.012, 95% CI [-0.003, 0.028], p = .111). This finding is perhaps unsurprising, given that—unlike the Facebook experiment—we only included users in the campaign if they had recently shared misinformation pre-treatment. Similarly, there was no significant difference in the size of the treatment between users classified as Democrats versus Republicans (interaction between treatment dummy and partisanship: b = -0.020, 95% CI [-0.072, 0.031], p = .439) based on the politicians and organizations they followed[36]; between users who followed any politicians/organizations versus none (interaction between treatment dummy and political follower dummy: b = -0.006, 95% CI [-0.053, 0.041], p = .814); or based on the users' number of followers (interaction between treatment dummy and log[followers+1]: b = -0.008, 95% CI [-0.019, 0.004], p = .200).

We also found no evidence that the treatment decayed over the course of the study (i.e., no significant interaction between treatment and campaign day: b = 0.00003, 95% CI [-0.002, 0.002], p = .982). And, as in the Facebook experiment, we find a non-significant effect of the treatment on users' overall sharing of non-misinformation posts (b = -0.004, 95% CI [-0.025, 0.018], p = .746; Fig. 4B), and this effect is significantly more positive than the effect on misinformation sharing (b = 0.034, 95% CI [0.002, 0.067], p = .040; although we note that there is no significant effect on the fraction of tweets that are low quality, see SI section 2.2.6 for details). Thus, the treatment effect is likely specific to misinformation rather than reducing sharing more generally.

Finally, we consider the fourth experiment that involved low-activity users who had not recently shared misinformation. In line with the results for users who had not recently shared misinformation in the Facebook experiment, this fourth Twitter experiment found no significant treatment effect (ITT b = 0.016, 95% CI [-0.006, 0.038], p = .154; ATT b = 0.026, 95% CI [-0.010, 0.063], p = .158), and the treatment effect size in this study was meaningfully different from the other three experiments that targeted recent misinformation sharers (Q[1] = 10.27, p =

.001; for several additional metrics providing clear evidence of treatment heterogeneity across experiments, see SI section S2.2.4; unsurprisingly, when pooling across all four Twitter experiments despite this heterogeneity, the overall effect is reduced: ITT b = -0.008, 95% CI [-0.025, 0.008], p = .317; ATT b = -0.013, 95% CI [-0.040, 0.014], p = .353). These results reinforce the importance of targeting the intervention at users who recently shared misinformation.

Discussion

Here we have shown in two large-scale field studies that accuracy prompts can successfully reduce misinformation sharing on social media platforms. This approach is particularly promising because it does not require any information about the specific posts being targeted—that is, it is content-neutral. Thus, accuracy prompts can help complement more traditional content-specific approaches (e.g., fact-checking and algorithmic identification of problematic content), reducing the spread of misinformation that has not yet been identified or shared in contexts where identification is impossible (e.g., encrypted messaging apps).

To our knowledge, these studies represent the first evidence of content-neutral interventions applied at scale to reduce misinformation sharing and the first investigation of interventions across multiple major social media platforms. The quantitative similarity of results across our studies conducted on Facebook and Twitter, despite the many differences in study implementation, lends extra credence to the conclusions.

The magnitudes of the effects we document here—reductions in misinformation sharing among recent misinformation sharers of 2.6% to 6.3%—are in line with expectations based on prior work. Broadly, other digital ad experiments in commercial, political, and public health messaging contexts have found effects of similar magnitudes[20-22]. Specific to accuracy prompts, survey experiments typically find 9-10% reductions in sharing intentions for false claims[12,16], and there are numerous reasons to expect the observed effect in the ad field experiments to be smaller than the survey experiments. For example, in the survey experiments, all participants receive the treatment, the outcome is measured with perfect precision (as the experimenters choose which fact-checked false claims to expose users to), respondents may not predict their own behavior accurately, and there is no treatment interference between participants. Conversely, in the field experiments, many users assigned to treatment do not actually receive the treatment (e.g., because they scroll past the ad without looking at it), classifiers or domain-level quality ratings must be relied upon as rough estimates of information quality (which add substantial measurement noise and thus depress observed effect sizes), and there is the potential for interference across conditions (e.g., if a treatment user does not share a given post, users in the control may be less likely to see—and thus share—it). Given these and related factors, the observed field effects compare favorably to expectations based on the survey experiments.

There are reasons to believe that, although small, effects of this magnitude can be practically relevant. For example, because of network effects, small changes at the level of the individual may lead to substantially larger impacts at the level of the system because reduced sharing reduces exposure—and thus sharing—by one's followers, which in turn reduces exposure and sharing of their followers, and so on[10]. Relatedly, because of ranking algorithms, reductions in

sharing, even among users with few followers, can reduce the reach of a given post by reducing engagement signals passed to the algorithm. Furthermore, given the "existence proof" evidence we provide for campaigns of this nature, future work can explore other delivery methods that are likely to generate larger effect sizes (such as interstitials/pop-ups that are more difficult to ignore, or banners across the top of the screen which are less fleeting).

A key conclusion from these studies is that for accuracy prompts to be effective, frequent prompting is required (as the effect of each individual prompt probably does not persist much beyond the browsing session in which it is delivered—future work should quantify exactly how long such effects last). An important feature of the design of our Twitter study was that we used a wide variety of different prompts to avoid habituation to the intervention (and accordingly, we did not find evidence that the treatment effect decayed over repeated deliveries). Such variation seems critical for long-term intervention success.

Another important finding across both studies is that the interventions were most effective when targeted at users who were likely to share misinformation in the absence of treatment. Treating users who would not have shared misinformation anyway is not cost-effective and even has the potential for perverse effects. This targeting of "at risk" users could be done at the level of the user (e.g., by identifying users who recently shared misinformation, although this would not work in the context of encrypted messaging apps), but also could be done by deploying the intervention at moments (e.g., natural disasters, elections, protests) when spikes in misinformation sharing are anticipated. Critically, though, the importance of targeting active misinformation sharers is not specific to accuracy prompts but applies to any interventions seeking to reduce the sharing of misinformation. That said, targeting may be viewed as discriminatory, which may pose challenges in many contexts - although one advantage of accuracy prompts is that they do not in any way restrict the actions of the targeted users but rather simply help them make choices that are informed by their own preferences. Additionally, even with a targeted application of the intervention, it would still be important to deploy cost-effectively to facilitate adoption by platforms (e.g., ads may not be the optimal medium for delivery) and ensure that the treatment did not lead to a reduction in non-misinformation sharing.

Finally, beyond insights into deploying accuracy prompts interventions at scale, our paper also makes important contributions to the understanding of online misinformation sharing more broadly. The data from our Facebook experiment offer the largest scale assessment to date of the prevalence of misinformation sharing on social media, examining numerous orders of magnitude more users than most prior work[28,29] and utilizing a misinformation measure that considers all posts rather than being limited to domains, URL content or political content as in past work[28,29,37] (see SI Section 1.2). Furthermore, our observational analysis of misinformation sharing before versus after identification by 3rd party fact-checkers (see SI Section 1.3) is one of the first to provide evidence regarding the actual on-platform impact of existing enforcement policies.

In sum, the data presented here suggest that accuracy prompts—when deployed appropriately—provide a promising content-neutral approach that may complement or even augment[38] existing content-specific interventions in reducing the spread of misinformation online.

Methods

Facebook experiment

***Campaign delivery and data collection*.** The Meta research team conducted the Facebook ad campaign targeting users in the United States from January 19 to February 10, 2023. The campaign audience consisted of a group of almost 33 million U.S. Facebook users selected without specific targeting criteria (representing a broad Facebook user population in the U.S., but not necessarily representative of U.S. Facebook users as a whole because the treatment could be crowded out or outbid by other content using the ads delivery infrastructure that is targeted to specific subgroups), as well as roughly 350,000 U.S. Facebook users who had recently shared misinformation (i.e., users who, as of 2 months prior to the campaign, had shared at least one post labeled as misinformation by fact-checkers in the 30 preceding days and at least 3 posts labeled as misinformation by fact-checkers in the 90 preceding days). See Table 1.

Users were assigned to the control or treatment group, and users in the treatment group were further randomized to receive one of three treatments that prompted accuracy in three different ways (SI section S1.1; Fig. S1): 1) Digital Literacy group: static images providing "tips for thinking critically before sharing," 2) Video group: a nine-second video stating that some stories use emotional language and encouraging users to "check for accuracy," or 3) Random Selection group: an ad randomly selected from the digital literacy tips images, a "poll ad" asking users how important accuracy is when sharing, or a message that said that 88% of Americans believe it is important that the news they read online is accurate (this statistic was obtained from a separate internal off-platform survey of U.S. adults conducted in August 2021); unfortunately, which specific ad was shown at which time for the Random Selection treatment was not logged. Prior survey experiments have demonstrated that these types of interventions are all similarly effective at shifting attention to accuracy[7].

Users in the treatment were shown accuracy prompt ads 3.2 times on average during the three-week campaign. Users in the control group were also assigned accuracy prompt ad spots but were shown standard Facebook ads rather than accuracy prompts.

Analysis. We compare the sharing of posts containing misinformation by users in the control versus treatment groups using linear regression. We evaluate all posts by users and classify a post as containing misinformation if it was flagged by third-party fact-checkers, a representative group of Facebook users (via Meta's Community Review team), or the predictions of an internal Facebook misinformation classification algorithm.

The main analysis focused on the 60-minute window immediately after an accuracy prompt treatment was delivered. Because the majority of users who shared any misinformation had shared only one misinformation post (i.e., 60.6% when examining baseline data from the week prior to the study), our main analyses used linear probability models to predict whether or not a given user shared any misinformation during this 60-minute window. For readability, we indicate all coefficients in units of basis points (i.e. multiply all coefficients by 10^5).

Table 1. Number of Facebook users by treatment group.

	Digital Literacy		Video		Random Selection	
Group	Control	Treatment	Control	Treatment	Control	Treatment
Non-Targeted Audience	5,313,712	5,305,307	5,266,611	5,256,049	5,781,630	5,768,251
Targeted cluster	58,394	59,156	58,764	58,289	59,279	58,729

Twitter study

Campaign delivery and data collection. The academic research team conducted the Twitter study, which consists of four different experiments conducted between October 2021 and April 2022 (Table 2). Each experiment included unique, non-overlapping Twitter users, and all data were collected using the official Twitter API. Experiments R1, R2, and R3 targeted active users who recently shared low-quality content; experiment NR targeted users who had not recently shared low-quality content, but had done so at some time in the past. The measures and analyses for Experiments R2, R3, and NR were pre-registered before data analysis. See Section S2.1.5 for pre-registration details.

Table 2. Twitter campaign descriptives by experiment and treatment group.

Experiment	Country	Group	Sample Size	Users Reached	Daily Impressions/User
R1 (8 days)	U.S.	Control	16,473	-	-
	U.S.	Treatment	16,415	8,424 (51.3%)	3.46
R2 (22 days)	Canada	Control	11,948	-	-
	Canada	Treatment	11,885	7,641 (64.3%)	2.18
R3 (9 days)	U.S.	Control	9,525	-	-
	U.S.	Treatment	9,517	5,952 (62.5%)	3.08
NR (12 days)	U.S.	Control	40,619	-	-
	U.S.	Treatment	40,671	25,135 (61.8%)	3.18

Experiment R1 was an 8-day Twitter campaign conducted between October 17 and 24, 2021. An 11-day baseline window, September 23 to October 3, 2021, was used to compute a pre-campaign covariate. 32,888 Twitter users were included in this experiment because they shared at least one of 533 lower-quality domains between September 23 and 30. We used block randomization[39] (R library quickblock [v0.2.0]) to assign users to either the control or treatment group (covariates used for randomization included 15 features such as users' number of friends, account age, number of tweets with lower-quality domains; see SI section S2.1.1 for full list of features). Of the 16,415 users assigned to the treatment group, 8,424 (51.32% "reach") saw one of our ads at

least once. The ads received 233,451 impressions in total or 29,181 per day, and on average, each user saw 3.46 ads per day.

Experiment R2 was a 22-day Twitter campaign conducted between February 22 and March 15, 2022, during the Canadian convoy protest. A 9-day baseline window, February 13 to 21, 2022, was used to compute a pre-campaign covariate. 23,833 Twitter users were included because during February 11 to 16, they posted content that included at least one of 34 hashtags (e.g., #canadatruckers, #freedomconvoy, #trudeauresign; see SI section S2.1.2 for all the hashtags used to identify eligible users). We used blocked randomization to assign users to either the control or treatment group (covariates for randomization included 10 features such as number of relevant hashtags shared, users' number of friends, account age; see SI section S2.1.1 for full list of features). Of the 11,885 users assigned to the treatment group, the campaign reached 7,641 users (64.29%). The ads received 366,918 impressions in total or 16,678 per day, and each user saw 2.18 ads per day.

Experiment R3 was a 9-day Twitter campaign conducted between April 19 and 27, 2022. A 13-day baseline window, April 6 to 18, 2022, was used to compute a pre-campaign covariate. We identified 19,042 Twitter users who shared at least one tweet containing the phrases "deepstate" or "deep state" between April 1 and 14, and used blocked randomization to assign them to the control or treatment group (covariates for randomization included 9 features such as users' number of friends, account age; see SI section S2.1.1 for full list of features). Of the 9,517 users assigned to the treatment group, the campaign reached 5,952 users (62.54%). The ads received 165,018 impressions in total or 18,335 per day, and each user saw 3.08 ads per day.

Experiment NR was a 12-day Twitter campaign conducted between October 22 and November 3, 2021. A 21-day baseline window, October 1 to 21, was used to compute a pre-campaign covariate. Because this experiment used the same user identification strategy as experiment R1 (which also overlapped with this experiment from October 22 to 24), we (i) excluded users targeted in Experiment R1 (which means excluding all users who had shared at least one of the lower-quality domain links 3 weeks prior to the experiment) and (ii) expanded the query window to ensure we could identify enough users: Any user who shared at least one low-quality domain since August 1 was eligible. This combination of excluding relatively recent misinformation sharers and including users who had shared misinformation up to 2.5 months earlier means that the users in this sample were not recent misinformation sharers and were much less active (see SI section S2.1.3). In total, 81,290 Twitter users were included in this experiment, and we used blocked randomization (same covariates as experiment R1) to assign users to the control or treatment group. Of the 40,671 users assigned to the treatment group, the campaign reached 25,135 users (61.80%). The ads received 958,883 impressions in total (79,907/day), and each user saw 3.18 ads per day.

Analysis. The main dependent variable for the three experiments with predominantly U.S. users is the number of "low quality" domains shared by users in their "retweets." We focus our analyses on 11,520 domains evaluated using a "wisdom of experts" approach in previous work[27]. As with previous work[27], we excluded popular non-news domains (e.g., google.com, youtube.com, facebook.com). The quality of domains ranges from 0 (lowest quality: naturalnews.com) to 1 (highest quality: reuters.com). For the main analyses, we classify a

retweet as a low-quality domain if it is at or below 0.70; in the SI section S2.2.2, we report analyses with domain quality thresholds ranging from 0.40 to 0.80 (in steps of 0.05) and show that our results are relatively robust to using different quality thresholds. The main dependent variable for the experiment with predominantly Canadian users is the total number of hashtags shared by users in their "retweets," because the users were selected based on sharing one of 34 hashtags (prior to the experiment) associated with an anti-vaccination protest in Canada, Ottawa. As pre-registered, after the experiment ended, we searched for the most popular and relevant hashtags shared during the protest using Twitter's API (rather than by looking at our data) and focused on these hashtags in our main analyses. In the SI section S2.2.3, we report analyses using two other dependent variables for this experiment: the number of retweets containing one of these hashtags and the domain-based approach described above.

For each experiment, we fit fixed-effects quasi-Poisson regression models to estimate the treatment effect. All models include two fixed effects: randomization blocks and campaign dates because in each experiment, blocked random assignment was used to assign users to the control and treatment groups (coded -0.5 and 0.5, respectively), and the dependent variables ("time1") were measured daily. Standard errors are clustered on randomization blocks. To increase precision, we include the corresponding pre-campaign "time0" variable in each model as a covariate. This "time0" covariate is sqrt-root transformed (to address right skew) and mean-centered prior to model fitting. We fit the models using the R (v4.3.1) library fixest (v0.11.1), using the following model specification: feglm(time1 ~ condition * time0 | block + date, cluster=~block, family="quasipoisson"). Both the "time1" and "time0" variables were separately winsorized by replacing values above the 95th percentile of all values with the 95th percentile of all values. We report the results of winsorizing values at the 99th percentile and no winsorizing in the SI section S2.2.7.

We conducted additional analyses using the same model specification—but the "time1" and "time0" variables are the number of non-low-quality retweets users shared—to investigate whether the treatment effect was specific to low-quality content rather than reducing sharing more generally.

We examined whether the treatment effect on the number of low-quality domains shared was moderated by other important covariates, including political affiliation (Republican or not) and whether users followed politicians or organizations or not, using the following fixed-effects quasi-Poisson: feglm(time1 ~ condition * (time0 + covariate) | block + date, cluster=~block, family="quasipoisson").

We assess the effect of the intervention by estimating the "intent-to-treat effect" (ITT) and the "average treatment effect on the treated" (ATT). The ITT effect is the average causal effect of treatment assignment; it represents the overall effect of the intervention, regardless of whether a user in the treatment group had actually seen the ads. The ATT effect is the effect of the treatment on the treated, and the estimand is the average effect among "compliers," the subset of users who would be treated if assigned to the treatment group. To estimate the ATT effects, separately for each experiment, we divide the ITT estimates by the proportion of users that saw the ads (0.513, 0.643, and 0.625 for the three experiments). The standard errors for the ATT effects were

computed via bootstrapping: For each model, we create 5,000 bootstrap samples by sampling with replacement at the level of the randomization block.

To estimate the average effect across the experiments, we aggregate the model coefficients from the different experiments by performing fixed-effects meta-analyses, also known as "common effects," because it assumes a single true effect exists that is common to all the observed studies.

Acknowledgments

The academic research team thanks Antonio Arechar, Adam Bear, Puneet Bhargava, Rocky Cole, Ziv Epstein, Beth Goldberg, Andrew Gully, and Mohsen Mosleh for helpful input, feedback, and comments on the design and execution of the Twitter experiments, Michael Stagnaro for feedback on the presentation of the results, Niko Lin for graphic design, and the Natural Sciences and Engineering Research Council of Canada, The Canadian Heritage Foundation, and The Government of Canada for providing funding.

Competing interests

Garro, Wernerfelt, Shore, Hughes, and Deisenroth were employees of Meta while the research was conducted. Berinsky and Pennycook were Faculty Research Fellows at Google in 2022. Berinsky was a paid consultant and Rand was an unpaid consultant for Twitter in 2021 and 2022. Eckles was a paid consultant for Twitter in 2022 and 2023. Other research by Berinsky, Eckles, Pennycook, and Rand has been funded by Meta and other work by Berinsky, Pennycook, and Rand has been funded by Google. Meta has sponsored a conference that Eckles organizes.

References

1. Lazer, D. M. J. et al. The science of fake news. *Science* **359**, 1094-1096 (2018).
2. Kozyreva, A., Lewandowsky, S. & Hertwig, R. Citizens versus the internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest* **21**, 103-156 (2020).
3. Martel, C., Allen, J., Pennycook, G. & Rand, D. G. Crowds can effectively identify misinformation at scale. *Perspectives on Psychological Science* 17456916231190388 (2023).
4. Martel, C. & Rand, D. Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology* 1-10 (2023).
5. DOMO. Data never sleeps 10.0 (internet archive). (2023).
6. Jaimungal, C. America speaks: Do they think fact-checks of political speeches are helpful? *YouGov* (2020).
7. Epstein, Z. et al. Developing an accuracy-prompt toolkit to reduce covid-19 misinformation online. *Harvard Kennedy School Misinformation Review* **2**, 1-12 (2021).
8. Guess, A. M. et al. 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* **117**, 15536-15545 (2020).
9. Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G. & Rand, D. G. Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science* **31**, 770-780 (2020).
10. Pennycook, G. et al. Shifting attention to accuracy can reduce misinformation online. *Nature* **592**, 590-595 (2021).
11. Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S. & Lewandowsky, S. Psychological inoculation improves resilience against misinformation on social media. *Science Advances* **8**, eabo6254 (2022).
12. Arechar, A. A. et al. Understanding and combatting misinformation across 16 countries on six continents. *Nature Human Behaviour* **7**, 1502-1513 (2023).

13. Epstein, Z., Sirlin, N., Arechar, A., Pennycook, G. & Rand, D. The social media context interferes with truth discernment. *Science Advances* **9**, eabo6169 (2023).
14. OECD. Misinformation and disinformation: An international effort using behavioural science to tackle the spread of misinformation. *OECD Public Governance Policy Papers* **21**, (2022).
15. Offer-Westort, M., Rosenzweig, L. R. & Athey, S. Battling the coronavirus `infodemic' among social media users in africa. *arXiv* 2212.13638v1 (2022).
16. Pennycook, G. & Rand, D. G. Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications* **13**, 1-12 (2022).
17. Stock, F., Hertwig, R. & Lorenz-Spreen, P. An accuracy self-nudge to reduce misinformation sharing online. *PsyArXiv* (2023).
18. Rathje, S., Roozenbeek, J., Traberg, C., Van Bavel, J. & van der Linden, S. Letter to the editors of psychological science: Meta-analysis reveals that accuracy nudges have little to no effect for u.s. Conservatives: Regarding pennycook et al. (2020). *Psychological Science* (2022).
19. Pretus, C. et al. The role of political devotion in sharing partisan misinformation and resistance to fact-checking. *Journal of Experimental Psychology: General* (2023).
20. Gordon, B. R., Moakler, R. & Zettelmeyer, F. Close enough? A large-scale exploration of non-experimental approaches to advertising measurement. *Marketing Science* **42**, 768-793 (2022).
21. Athey, S., Grabarz, K., Luca, M. & Wernerfelt, N. 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* **120**, e2208110120 (2023).
22. Aggarwal, M. et al. A 2 million-person, campaign-wide field experiment shows how digital advertising affects voter turnout. *Nature Human Behaviour* **7**, 332-341 (2023).
23. Silverman, H. Helping fact-checkers identify false claims faster. *Meta* (2019).
24. Meta. How fact-checking works. *Meta Transparency Center* (2022).
25. Lasser, J. et al. Social media sharing of low-quality news sources by political elites. *PNAS Nexus* **2**, pgad158 (2022).
26. Lasser, J. et al. From alternative conceptions of honesty to alternative facts in communications by us politicians. *Nature Human Behaviour* 1-12 (2023).
27. Lin, H. et al. High level of correspondence across different news domain quality rating sets. *PNAS Nexus* **2**, 1-8 (2023).
28. Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. & Lazer, D. Fake news on Twitter during the 2016 U.S. Presidential Election. *Science* **363**, 374-378 (2019).
29. Guess, A., Nagler, J. & Tucker, J. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances* **5**, eaau4586 (2019).
30. Vigderman, A. 90% of people claim they fact-check news stories as trust in media plummets (internet archive). (2023).
31. Guess, A. M. Measure for measure: An experimental test of online political media exposure. *Political Analysis* **23**, 59-75 (2015).
32. Prior, M. The immensely inflated news audience: Assessing bias in self-reported news exposure. *Public Opinion Quarterly* **73**, 130-143 (2009).
33. Guay, B., Berinsky, A. J., Pennycook, G. & Rand, D. How to think about whether misinformation interventions work. *Nature Human Behaviour* **7**, 1231-1233 (2023).
34. Madore, K. P. et al. Memory failure predicted by attention lapsing and media multitasking. *Nature* **587**, 87-91 (2020).
35. Pearson, D., Watson, P., Albertella, L. & Le Pelley, M. E. Attentional economics links value-modulated attentional capture and decision-making. *Nature Reviews Psychology* **1**, 320-333 (2022).
36. Mosleh, M. & Rand, D. G. Measuring exposure to misinformation from political elites on twitter. *Nature Communications* **13**, 7144 (2022).
37. González-Bailón, S. et al. Asymmetric ideological segregation in exposure to political news on Facebook. *Science* **381**, 392-398 (2023).
38. Pennycook, G. et al. Misinformation inoculations must be boosted by accuracy prompts to improve judgments of truth. *PsyArXiv* (2023).
39. Higgins, M. J., Sävje, F. & Sekhon, J. S. Improving massive experiments with threshold blocking. *Proceedings of the National Academy of Sciences* **113**, 7369-7376 (2016).

Supplementary Information for

Reducing misinformation sharing at scale using digital accuracy prompt ads

Hause Lin[1], Haritz Garro[1], Nils Wernerfelt, Jesse Shore, Adam Hughes, Daniel Deisenroth, Nathaniel Barr, Adam Berinsky, Dean Eckles, Gordon Pennycook*, & David G. Rand*

[1]These authors contributed equally: H. Lin, H. Garro.
*Corresponding authors: gordon.pennycook@cornell.edu, drand@mit.edu

Contents

S1 Facebook Study

S1.1 Ad Creatives

Figure S1. Facebook ad creatives shown to users in the treatment group. *(**A**) Users in the Digital Literacy group saw static images that provided tips for thinking critically before sharing content. (**B**) Users in the Video group were shown a short nine-second video stressing that some stories may use strong, emotional language to evoke a strong reaction and encouraged users to check for accuracy. (**C**) Users in the Random Selection group were shown different messages that included the three digital literacy creatives, a "poll" ad asking them how important accuracy is when sharing a post or article, and a message that said "88% of Americans believe that it is important that the news they read online is accurate." Internal Meta designers worked with an external agency to ensure these creatives followed best practices, and relied upon previous creatives that the Global Policy Programs team had launched in their informational campaigns to fight against COVID-19 misinformation. All ads included the Meta branding.*

S1.2 Baseline Descriptives

To provide a sense of the baseline for our key outcome of misinformation sharing, we provide a description of behavior in the week prior to the beginning of our experiment. Of the 33 million users in our experiment, 2.13 million shared at least one piece of likely misinformation (i.e., flagged by either 3rd party fact-checkers (3PFC), Community Review crowd raters, or the misinformation classifier) in the week before the experiment. Out of these, 1.866 million users (87.6%) were from the non-targeted audience and 263,000 users (12.4%) were from the users targeted because they had repeatedly shared links classified as misinformation prior to the study. The distribution of the number of pieces of likely misinformation shared among sharers is very right-skewed (Figure S2). For example, 60.6% of the users who shared any likely misinformation in the pre-experiment week (1.29 million users) only shared 1 piece of likely misinformation.

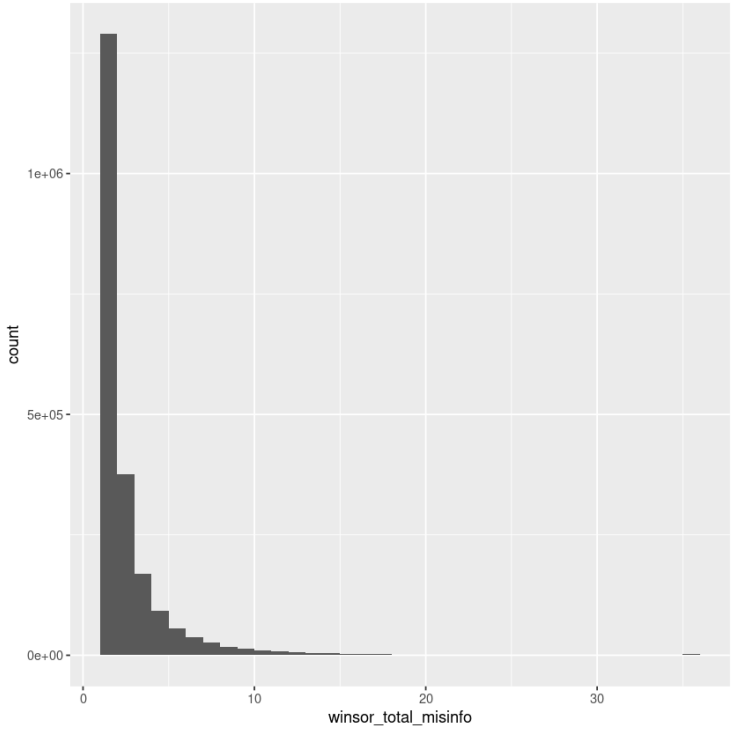

Figure S2. *Distribution of the number of pieces of likely misinformation shared in the pre-experiment week. The x-axis is clipped at the 99.9th percentile (i.e., users that shared more than 36 pieces of misinformation).*

S1.3 Baseline Standard of Care

The effect of the content-free intervention we examine here must be understood in the context of existing content-specific interventions that are currently deployed on platform (and with which our intervention was thus implemented in tandem). To provide some insight into the baseline (content-specific) standard of care, we present an observational analysis of sharing before versus after a piece of content is identified as misinformation by 3PFCs (which leads to labeling and warning if the user attempts to reshare, and may be subject to demotion). To do so, we examine all posts reshared by users in the targeted cluster during our 3 week experiment that were identified as 3PFC misinformation. (Slightly less than 50% of all 3PFC misinformation reshared during the campaign period was identified as misinformation before, during, or shortly after the campaign, and it is these posts that we examine here.) For each sharing event, we calculate the distance (in days) between when the sharing event occurred and when the post was identified as 3PFC misinformation (negative values indicate that the share occurred before the content was identified as 3PFC, positive values indicate that the share occurred after the content was identified as 3PFC). A histogram of the resulting values is shown in Figure S3. Although the observational nature of these data obviously do not allow strong causal inferences, the pattern is striking: Once Meta identifies content as 3PFC misinformation and applies the baseline standard of care, resharing is almost entirely eliminated.

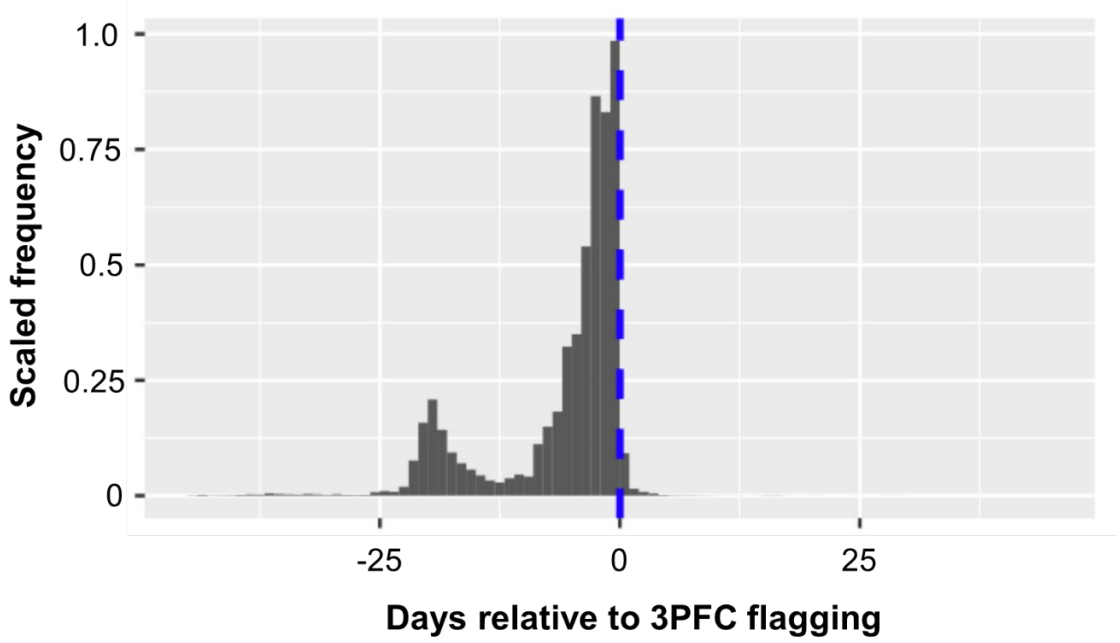

Figure S3. *Histogram of number of days between Meta identifying a post as 3PFC misinformation and that post being reshared. Negative values indicate that the reshare occurred before the content was identified as 3PFC; positive values indicate that the reshare occurred after the content was identified as 3PFC. Y-axis is scaled such that the largest bin is 1.*

S1.4 Results

Here we present regression tables for the analyses described in the main text, along with alternative specifications that use quasi-Poisson models to predict the number of misinformation posts shared (rather than OLS models predicting whether users shared at least 1 misinformation post) and which consider all users versus just users who shared at least one misinformation post in the pre-experiment week.

Table S1. *Models predicting misinformation sharing across all users in the 60 minutes after prompt delivery and across the full 3 weeks of the campaign.*

	1	2	3	4
Outcome	>0 Misinfo Post	# Misinfo Post	>0 Misinfo Post	# Misinfo Post
Model	OLS	Quasi-Poisson	OLS	Quasi-Poisson
Users	All	All	All	All
Time range	60m	60m	Full Campaign	Full Campaign
Treatment b	-0.000038	-0.0200	-0.00012	-0.0015
se	0.000016	0.0086	0.00012	0.0021
p	p = 0.018	p = 0.020	p = 0.313	p = 0.477
Intercept	0.002125	-6.0261	0.1344	-0.7123
Sample Size	33043471	33043471	33043471	33043471

Table S2. *Models examining moderation of the treatment effect based on having shared misinformation in the pre-experiment week, for the 60 minutes after treatment and the full campaign.*

	1	2	3	4
Outcome	>0 Misinfo Post	# Misinfo Post	>0 Misinfo Post	# Misinfo Post
Model	OLS	Quasi-Poisson	OLS	Quasi-Poisson
Users	All	All	All	All
Time range	60m	60m	Full Campaign	Full Campaign
Treatment b	-0.00057	-0.0296	-0.0002	-0.002
se	0.0002	0.0096	0.0002	0.0015
p	p = 0.004	p = 0.002	p = 0.212	p = 0.169
(0 pre-exp shares) b	-0.02113	-3.475	-0.6315	-3.2836
se	0.00014	0.012	0.0001	0.0017
p	p = 0.000	p = 0.000	p = 0.000	p = 0.000
Treatment X (0 pre-exp shares) b	0.00057	0.0292	0.0001	0.0001
se	0.0002	0.0171	0.0002	0.0024
p	p = 0.004	p = 0.087	p = 0.604	p = 0.969
Intercept	0.0219	-3.6547	0.7253	1.5956
Sample Size	33043471	33043471	33043471	33043471

Table S3. *Models examining each of the three treatment arms separately.*

	1	2	3	4
Outcome	>0 Misinfo Post	# Misinfo Post	>0 Misinfo Post	# Misinfo Post
Model	OLS	Quasi-Poisson	OLS	Quasi-Poisson
Users	>0 misinfo pre-exp shares		All	
Time range	60m	60m	60m	60m
Tips b	-0.00022	-0.0137	-0.000028	-0.0151
se	0.00029	0.0151	0.000023	0.0122
p	0.431	0.366	0.216	0.218
Video b	-0.00074	-0.0321	-0.00003	-0.0113
se	0.00028	0.0153	0.000023	0.0123
p	0.009	0.036	0.186	0.355
Combo b	-0.00074	-0.0423	-0.000054	-0.0324
se	0.00027	0.0148	0.000022	0.0119
p	0.007	0.004	0.015	0.007
Sample Size	2128311	2128311	33043471	33043471

Table S4. *Models examining moderation of the treatment effect by individual differences.*

	1	2	3	4
Outcome	>0 Misinfo Post	# Misinfo Post	>0 Misinfo Post	# Misinfo Post
Model	OLS	Quasi-Poisson	OLS	Quasi-Poisson
Users	>0 misinfo shares pre-exp		All	
Time range	60m	60m	60m	60m
Treatment b	-0.0004	-0.0147	-0.000031	-0.0136
se	0.0004	0.0179	0.000026	0.0141
p	0.235	0.413	0.232	0.335
female b	-0.0062	-0.3228	0.0000004	-0.0457
se	0.0003	0.0149	0.000023	0.012
p	0.000	0.000	0.988	0.0001
female X Treatment b	0.0002	-0.0093	0.000011	0.0022
se	0.0004	0.0212	0.000032	0.0171
p	0.688	0.662	0.741	0.898
Over65 b	0.0175	0.7442	0.00365	1.1925
se	0.0005	0.0166	0.000066	0.0144
p	0.000	0.000	0.000	0.000
Over65 X Treatment b	-0.0008	-0.0157	-0.00013	-0.0177
se	0.0007	0.0237	0.00009	0.0205
p	0.264	0.507	0.166	0.388
College b	-0.00001	-0.0057	-0.00035	-0.1601
se	0.00029	0.0153	0.00002	0.0124
p	0.963	0.711	0.000	0.000
College X Treatment b	-0.0002	-0.0098	0.0000038	-0.0043
se	0.0004	0.0219	0.0000324	0.0177
p	0.66	0.654	0.908	0.809
Intercept	0.0226	-3.6474	0.002	-6.1112
Sample Size	2108334	2108334	32740024	32740024

Table S5. *Models examining treatment effect on sharing of non-misinformation posts.*

	1	2
Outcome	# Total Non-Misinfo Posts	# Total Non-Misinfo Posts
Model	Quasi-Poisson	Quasi-Poisson
Users	All	>0 misinfo shares pre-exp
Time range	60m	60m
Treatment b	-0.0047	-0.0059
se	0.0028	0.0044
p	0.090	0.182
Sample Size	33043471	2128311

Finally, to facilitate reproducibility, we provide descriptive tables that contain the information required to reproduce the main analyses (i.e. the OLS models in Tables S1, S2 and S5) in Table S6.

Table S6. *Descriptive statistics for reproducing the main analyses.*

Condition (0=control, 1=treatment)	Audience (0=untargeted, 1=targeted)	Shared Misinfo in Pre-Exp Week	Shared Anything	Shared Any Misinfo	User Count	Time Period
0	0	0	0	0	14819402	60m
0	0	0	1	0	598230	60m
0	0	0	1	1	10897	60m
0	0	1	0	0	708492	60m
0	0	1	1	0	211219	60m
0	0	1	1	1	13713	60m
0	1	0	0	0	33762	60m
0	1	0	1	0	10007	60m
0	1	0	1	1	934	60m
0	1	1	0	0	73906	60m
0	1	1	1	0	47846	60m
0	1	1	1	1	9605	60m
1	0	0	0	0	14787961	60m
1	0	0	1	0	598630	60m
1	0	0	1	1	10784	60m
1	0	1	0	0	708279	60m
1	0	1	1	0	210734	60m
1	0	1	1	1	13219	60m
1	1	0	0	0	33746	60m
1	1	0	1	0	9819	60m
1	1	0	1	1	988	60m
1	1	1	0	0	74426	60m
1	1	1	1	0	47407	60m
1	1	1	1	1	9465	60m
0	0	0	Not measured	0	14010040	Full campaign
0	0	0	Not measured	1	1418489	Full campaign
0	0	1	Not measured	0	286095	Full campaign
0	0	1	Not measured	1	647329	Full campaign
0	1	0	Not measured	0	12208	Full campaign
0	1	0	Not measured	1	32495	Full campaign
0	1	1	Not measured	0	6390	Full campaign
0	1	1	Not measured	1	124967	Full campaign
1	0	0	Not measured	0	13983850	Full campaign
1	0	0	Not measured	1	1413525	Full campaign
1	0	1	Not measured	0	285954	Full campaign
1	0	1	Not measured	1	646278	Full campaign
1	1	0	Not measured	0	12273	Full campaign
1	1	0	Not measured	1	32280	Full campaign
1	1	1	Not measured	0	6452	Full campaign
1	1	1	Not measured	1	124846	Full campaign

S1.5 Brand Lift Study

As part of the campaign, we also administered surveys to users in the control and treatment groups using the brand lift infrastructure. Each user was shown only a single question. The brand lift collects responses from both users who saw the accuracy prompt ads (i.e., treated users) as well as users who did not see the accuracy prompt ads (i.e., control users).

We asked the following two questions while the campaign was active:

Meta Cares About Users (CAU) question:

- Please agree or disagree with the following statement: Meta (formerly Facebook, Inc.) cares about its users.
    - Strongly agree
    - Agree
    - Neither agree nor disagree
    - Disagree
    - Strongly disagree

Meta Misinfo Success question:

- How successful is Meta (formerly Facebook, Inc.) at reducing false information?
    - Extremely successful
    - Quite successful
    - Somewhat successful
    - Not at all successful
    - I'm not sure

We also asked the following question once the campaign had ended:

- In the last 14 days, did you think about the accuracy of the posts that you read on Facebook?
    - Yes/No/I don't remember

***Table S7a**. Number of Facebook users who responded to each question per treatment group.*

	Digital Literacy				Video				Combination			
	Untargeted		Targeted		Untargeted		Targeted		Untargeted		Targeted	
	T	C	T	C	T	C	T	C	T	C	T	C
Meta CAU	1,936	2,001	2,013	1,874	1,918	1,999	1,857	2,006	1,936	1,986	1,846	2,012
Meta Success	1,929	2,000	1,568	1,994	1,900	1,992	1,479	1,995	1,923	1,988	1,563	1,991
14 days action	2,004	1,992	1,487	1,507	1,965	2,003	1,515	1,491	2,002	1,993	1,424	1,418

Table S7a shows the number of users who answered each question (brand lift surveys on Facebook have a roughly 2% response rate). Table S7b below shows the point estimates from the "Advanced lift" results, which are calibrated with other lift tests and account for the representativeness of the users who responded to the survey invitation. Because of this calibration, it is unfortunately not possible to pool the brand lift results across treatment arms or samples. Nonetheless, we see a consistent pattern across treatment arms and samples whereby the treatments increasing the fraction of users who recalled thinking about accuracy while on Facebook; and do not see large or consistent effects on attitudes towards Meta.

We also note that in terms of baseline levels in the control, 38 to 42 percent of the untargeted audience—and 44 to 48 percent of the targeted audience—responded Yes to the question about considering accuracy; 12 to 14 percent of the untargeted audience—and 5 to 6 percent of the targeted audience—responded "Agree" or "Strongly Agree" to the Meta CAU question; and 12 to 13 percent of the untargeted audience—and 5 percent of the targeted audience—responded "Quite Successful" or "Extremely Successful" to the Meta Misinfo Success question.

***Table S7b**. Point estimates from the "Advanced lift" results.*

Brand Lift Results	Meta CAU	Meta Misinfo Success	Recall Thinking about Accuracy
Digital Literacy Untargeted Audience	+0.3 pts	+0.6 pts	+1.2 pts
Video Untargeted Audience	-0.8 pts	-1.3 pts	+1.6 pts
Random Selection Untargeted Audience	-1.3 pts	-0.8 pts	+2.8 pts
Digital Literacy Targeted Audience	+0.3 pts	-0.3 pts	+1.3 pts
Video Targeted Audience	+0.1 pts	-0.3 pts	+2.4 pts
Random Selection Targeted Audience	-1.3 pts	-0.1 pts	+1.9 pts

S2 Twitter Study

S2.1 Extended Methods and Descriptives

S2.1.1 Covariates For Block Randomization Treatment Assignment

In each experiment, we used blocked randomization to assign users to either the control or treatment group to ensure balanced treatment assignment across groups and covariate levels. Experiments R1 and NR used 15 covariates: pretreatment quality mean, pretreatment quality count, pretreatment quality sum, number of active days, number of relevant retweets retrieved, number of relevant tweets retrieved, follower count, friend count, favorites count, overall tweet count, friend-follow ratio, days since account creation, predicted quality mean, predicted quality count, predicted quality sum. Experiment R2 used 10 covariates: number of relevant hashtags, number of active days, number of relevant retweets retrieved, number of relevant tweets retrieved, follower count, friend count, favorites count, overall tweet count, friend-follow ratio, days since account creation. Experiment R3 used 9 covariates: number of active days, number of relevant retweets retrieved, number of relevant tweets retrieved, follower count, friend count, favorites count, overall tweet count, friend-follow ratio, days since account creation.

S2.1.2 Hashtags For Identifying Eligible Users in Experiment R2

For Experiment R2, we included users who were participating in and sharing content on Twitter related to the Canadian anti-vaccination "trucker convoy" protest. From February 11 to 16, 2022, we identified 34 popular hashtags and then used these hashtags to search for users that would be eligible for our experiment: #canadatruckers #canadiantruckers #convoyforfreedom2022 #endthemandates #freedomconv #freedomconvoy #freedomconvoycanada2022 #freedomrally #freedomtruckers #freetamaralich #freetruckers #holdthelinecanada #honkhonk #honkhonk2022 #honkhonkhonk #nojabs #ottawaoccupied #ottawasiege #truckerconvoy #truckerconvoy2022 #truckersconvoy #truckersforfreedom #truckersforfreedom2022 #truckersforfreedomconvoy #truckersforfreedomconvoy2022 #truckyoutrudeau #trudeaudictatorshipmustgo #trudeauhasgottogo #trudeauisacoward #trudeaumustresign #trudeaumustresign2022 #trudeaunationaldisgrace #trudeauresign #truenorthstrongandfree. We used a broad range of hashtags to cover engagement with the protest movement. Much of the language in the hashtags relates to "truckers," but the protest was motivated by opposition to COVID-19 vaccinations. For further details see[1-3].

S2.1.3 User Characteristics

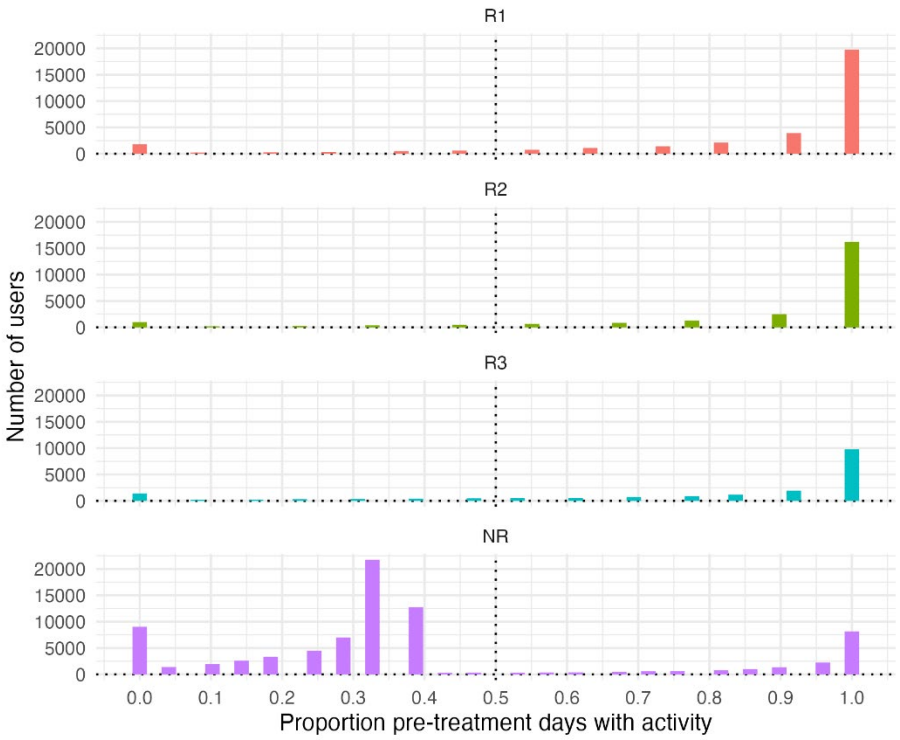

Figure S4. *Proportion of pre-treatment days with activity for each Twitter experiment. Experiment NR users were much less active before the experiment: It excluded relatively recent misinformation sharers and included users who had shared misinformation up to 2.5 months earlier.*

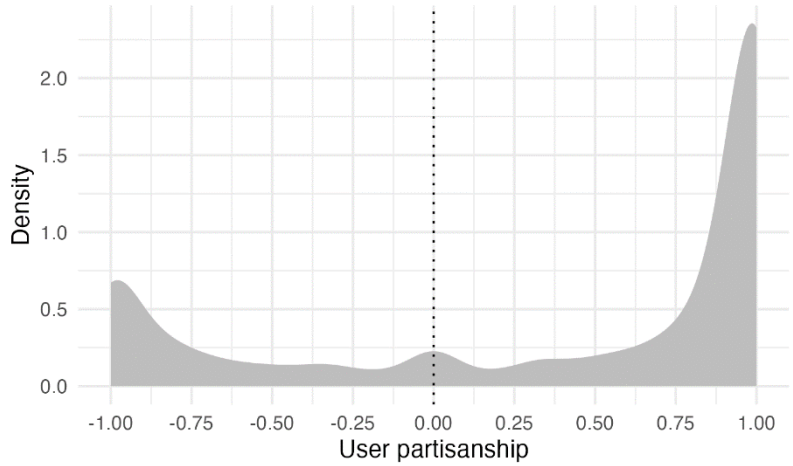

Figure S5. *User partisanship across all experiments. There were about three times as many Republican users as Democrat users. Partisanship was determined based on the politicians and organizations a user followed on Twitter.*

S2.1.4 Ad Creatives

The ad campaigns used a diverse set of 50 different accuracy prompt creatives, and the campaigns were delivered using the academic team's Twitter thinkaccuracy account (see profile page for example video creatives in the timeline). More examples are provided below. All creatives were videos that did not have accompanying audio.

Figure S6. *Four example video creatives used in the Twitter experiments. See @thinkaccuracy account's profile page for more example video creatives in the timeline.*

S2.1.5 Pre-Registrations

Experiment R1 was not pre-registered.

For Experiment R2, before data collection we submitted a pre-registration (https://osf.io/bqtmd). The experiment was run in immediate response to an unexpected unfolding crisis in Canada involving a large-scale protest against COVID-19 policies (e.g. mandates), and thus the pre-registration was assembled quickly and was not that detailed. The pre-registration specified the target sample (20k Twitter users as a goal, while acknowledging that the ultimate exact number would depend on Twitter's algorithms) and the basic model structure (condition, pre-experiment outcome, and their interaction as the independent variables), but not the specific modeling approach. In terms of outcomes, the pre-registration clearly specified that the number of convoy-related hashtags would be an outcome variable, with the set of hashtags to be considered determined by searching for the most popular hashtags using Twitter's API (rather than by looking in the datasets). We also pre-registered that we would create continuous ratings of domain quality including all links to domains that we have ratings for, although we did not specify what domain rating set we would use, or how we could calculate the quality ratings. A year after conducting the initial experiment—but prior to conducting any domain-level analyses—we submitted another pre-registration to more precisely specify how we could analyze domain quality (https://aspredicted.org/16g_kmz), using the same analyses described below in detail for Experiment R3. Given that the hashtag counting outcome was the only outcome that was clearly specified in the original pre-registration (and it has the advantage of being more precise than the domain-level approach), we focus on the count of relevant hashtags as our key outcome variable for study R2, and followed the pre-registered approach (as described in more detail in SI section S2.1.2). We then include the domain-level analyses in the second pre-registration as secondary measures in SI section S2.2.3, S2.2.5 and S2.2.6. Finally, the initial pre-registration also said that as a secondary analysis we would analyze primary tweets (rather than retweets) but there were comparatively few primary tweets with relevant hashtags (only 3.7% as many primary tweets as retweets) and thus we did not conduct these analyses due to lack of power; the initial pre-registration said that we would also fit models predicting post-campaign behavior, but did not specify anything about what this analysis would entail or how the outcomes would be calculated (e.g., how long after the end of the campaign we would examine) and indicated that we would count the number of "misinformation claims and/or low-quality content/domains" shared as an outcome, with the selection of domains and claims to be determined post hoc based on fact-checking sites and fact-checkers (but not specifying how precisely this would be done). Thus, any such analyses would be considered exploratory anyway, and given the large number of analyses in the paper, we did not analyze post-campaign behavior or identify claims based on post hoc fact-checking.

For Experiment R3, we submitted a pre-registration after data was collected, but before any data was analyzed (https://aspredicted.org/8ry_485). Below we reproduce the relevant section of the pre-registration, in which we describe how we will calculate the count of low quality domains shared (used as our primary outcome), a summed low quality score (see SI section S2.2.5), and the fraction of low-quality domains (see SI section S2.2.6):

> We have quality ratings for over 11,000 domains (obtained from a separate project). These are aggregated (via principal component analysis) ratings ("PC1"), which range from 0 (lowest quality) to 100 (highest quality). Domains with quality ratings below a particular threshold will be considered "low quality." We will run the analyses with thresholds ranging from 40 to 80, in steps of 5.
>
> Count of low-quality domains shared: Domains with ratings below the threshold will be considered low quality and counted (those above the threshold will be excluded because they are considered high quality). For each user, we will count the number of low-quality domains retweets, both before the campaign ("time 0") and during the campaign ("time 1"). Blocked random assignment was used to assign users to the control and treatment group, and the DVs/outcomes ("t1") were measured daily. Thus, we will fit fixed-effects quasipoisson regression models with blocks and days as fixed effects. Standard errors will be clustered on blocks. We will include the corresponding pre-campaign "t0" variable in the model as a covariate. Model specification/R syntax: fixest::feglm(t1 ~ condition * t0 | block + day, family = "quasipoisson", cluster = "block")
>
> Summed low-quality domain score: Domains with ratings below the threshold will be assigned the value 100 (maximum low quality), whereas domains with ratings above the threshold will be linearly rescaled to 99-0 (decreasing quality). The model specification is identical to the count model above: fixest::feglm(t1 ~ condition * t0 | block + day, family = "quasipoisson", cluster = "block")
>
> Fraction of low-quality domains shared: To compute this metric, we will divide the count of low-quality domains shared by the total number of retweets. Unlike summed badness and count of low-quality domains (which have no upper bound), this metric is bounded between 0 and 1, so we will fit linear regression models. Model specification/R syntax: fixest::feols(t1 ~ condition * t0 | block + day, cluster = "block")
>
> Probing significant interactions: If there are significant interactions, we will split the users into 5 bins (based on the covariate "t0") and then examine how the treatment effect differs across bins.
>
> Extreme and missing user values: We will winsorize all variables such that values above the 95th percentile will be replaced with the 95th percentile. Users with missing values will be assigned the value 0, reflecting that these users did not share any low-quality domains (note that this step happens after winsorizing the variables [previous step]).

For Experiment NR, we submitted a pre-registration after data was collected, but before any data was analyzed (https://aspredicted.org/mgf_mfr). The pre-registration was substantively identical to the pre-registration for Experiment R3, except that the summed low quality score was pre-registered as the primary outcome. For comparability across experiments, we instead use the count of low quality domains as the outcome in our main analysis, but the results are equivalent when using the summed low quality score (no significant treatment effect; see SI section S2.2.5)

Finally, in terms of aggregating the results of the four experiments, our decision to meta-analyze the results, and our choice of meta-analytic strategy, was determined after all data collection was completed and we had decided not to run any more experiments. Thus, the total number of experiments we conducted was not influenced by the results of meta-analysis.

S2.2 Results

S2.2.1 Individual Experiments

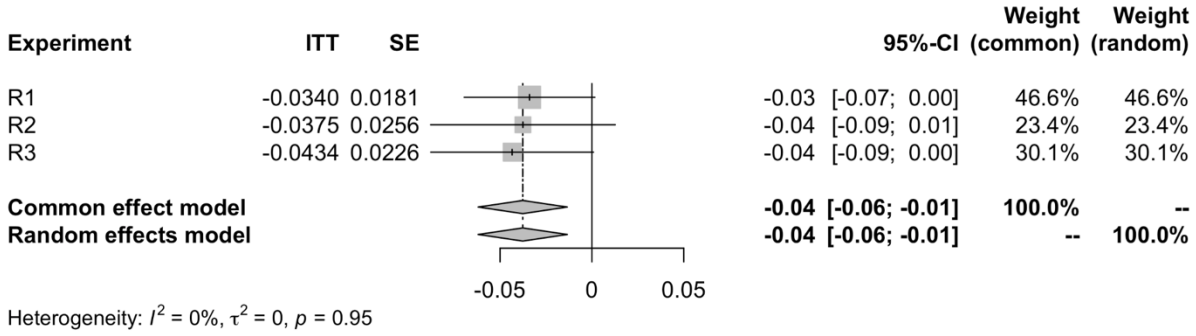

Figure S7*. Treatment effect (ITT) on reduction in misinformation sharing for each study and meta-analytic effect.*

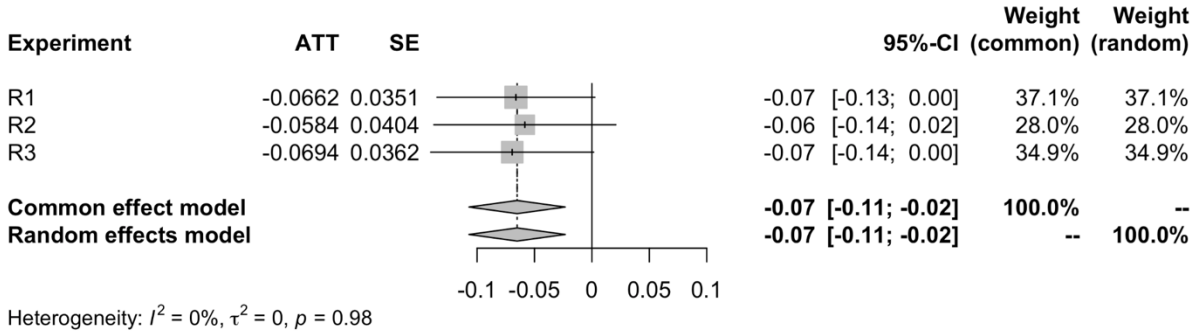

Figure S8*. Treatment effect on the treated (ATT) on reduction in misinformation sharing and meta-analytic effect.*

S2.2.2 Different Domain Quality Thresholds

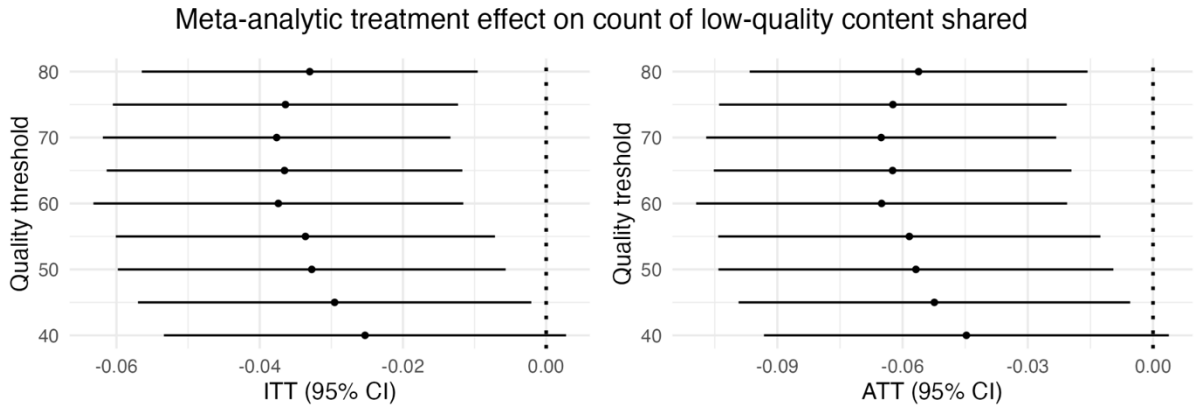

Figure S9*. Meta-analytic treatment effect for different domain quality thresholds. Analyses with higher thresholds included more domains but also included more higher quality domains. Analyses with lower thresholds included fewer domains but are more likely to contain only lower quality domains.*

S2.2.3 Other Dependent Measures for Experiment R2

For the analyses in the main text, Experiments R1 and R3 counted the number of low quality domains shared whereas Experiment R2 counted the number of relevant hashtags shared. We show that we also find significant meta-analytic treatment effects when using two alternative dependent measures for Experiment R2: using the *count of posts (i.e., retweets) with relevant misinformation hashtags* (which, unlike the number of hashtags, does not capture the strength of the signal given by the hashtags) gives ITT b = -0.034 [-0.058, -0.010], p = .005; ATT b = -0.059 [-0.101, -0.018], p = .005; using the number of low quality domains shared (which is a much coarser measure than hashtag count, as it is not specific to the content of posts) yields ITT b = -0.024 [-0.045, -0.003], p = .023; ATT b = -0.040 [-0.075, -0.004], p = .028.

S2.2.4 Treatment Effect Heterogeneity

There was no evidence of treatment effect heterogeneity across Experiments R1, R2, and R3 (ps > .95; see Figures S7, S8). Unsurprisingly—and as described in the main text—including Experiment NR in the meta-analysis led to significant heterogeneity (Q[3] = 10.37, p = .016; see Figures S10 and S11 below), because this experiment targeted relatively inactive users who had not recently shared links to low quality new sites, but who had done so further in the past. As expected, a meta-regression moderator analysis (predicting treatment effect using a Experiment NR dummy) showed that the difference between Experiment NR and the other experiments is highly significant (Q[1] = 10.27, p = .001; Figures S10 and S11). Importantly, this dummy

moderator in the meta-regression model accounted for 100% of the heterogeneity in treatment effects across experiments. Additional analyses using numerous different influence diagnostics (Figure S12) provides further evidence that Experiment NR is different from the other experiments (i.e., is an "influential case" that has a large effect on the pooled effect and heterogeneity), and that omitting Experiment NR results in a more precise estimate of the overall treatment effect.

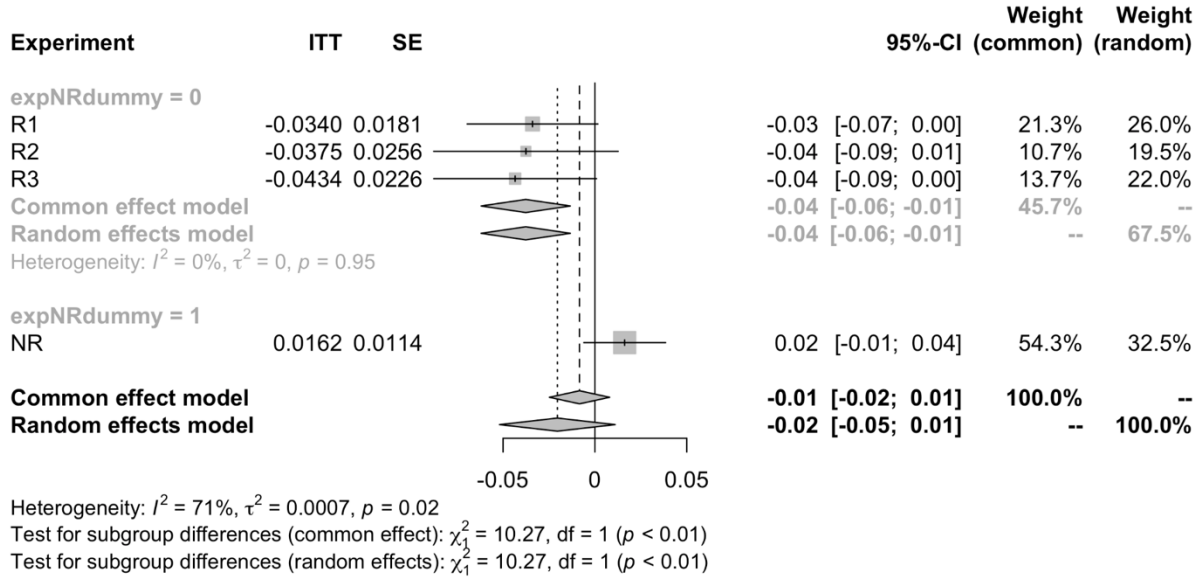

Figure S10*. Treatment effect (ITT) on reduction in misinformation sharing for each study and meta-analytic effect. Experiment NR targeted relatively inactive users who had not recently shared links to low quality new sites, but who had done so further in the past.*

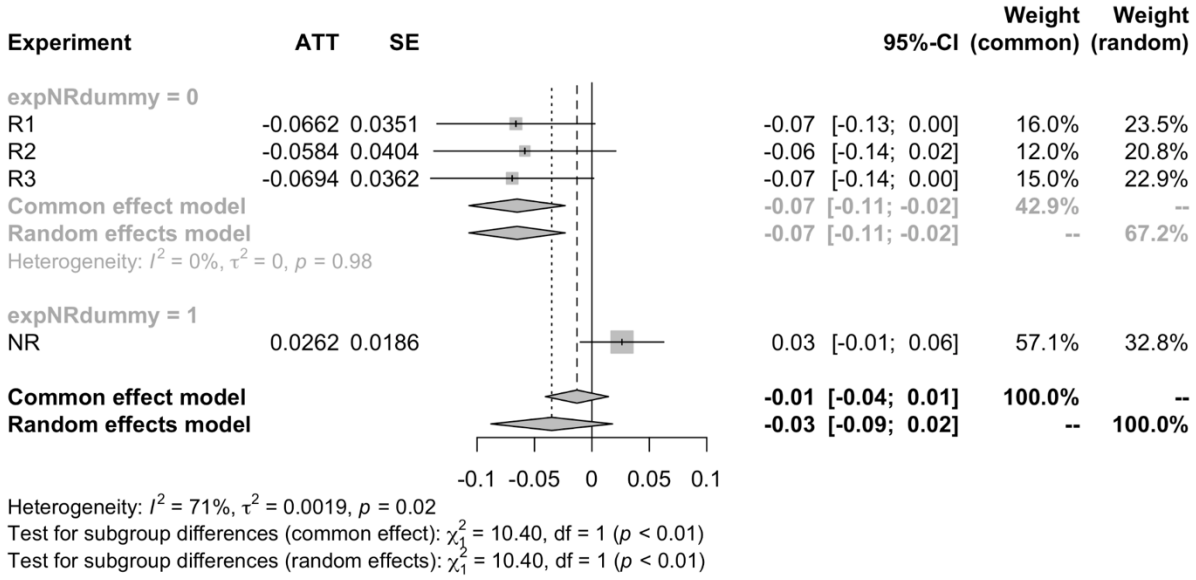

Figure S11*. Treatment effect (ATT) on reduction in misinformation sharing and meta-analytic effect. Experiment NR targeted relatively inactive users who had not recently shared links to low quality new sites, but who had done so further in the past.*

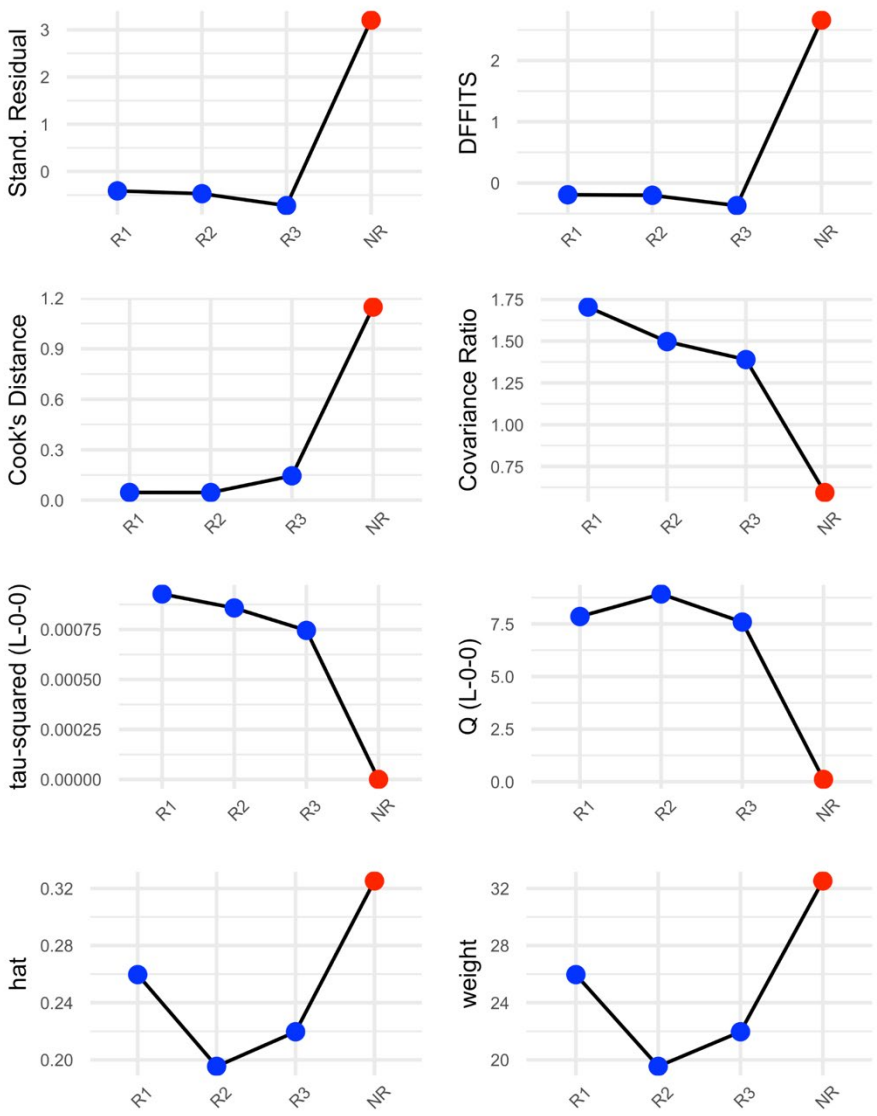

Figure S12. *Meta-analysis influence diagnostics for each experiment. Each panel is one diagnostic metric. Influence cases are shown in red. Higher values of the following metrics suggest an influential case: externally standardized residual (Stand. Residual), difference in fits (DFFITS), Cook's distance, hat (another measure of study weight), and weight. Lower values of the following metrics suggest an influential case: covariance ratio, leave-one-out tau-squared, and leave-one-out Q.*

S2.2.5 Graded Approach to Scoring Domain Quality

We evaluate the robustness of our results by using a graded approach to scoring domain quality. Domains with ratings below a given threshold (e.g., 0.70) are assigned the value 1.0 (maximally low quality), whereas domains with ratings above that threshold are linearly rescaled to 0.99 to 0. As with the main analyses (threshold = 0.70), users in the treatment shared less low quality

content than users in the control (ITT b = -0.028 [-0.051, -0.006], p = .014; ATT b = -0.047 [-0.086, -0.009], p = .017; see Figures S13 and S14). The results are robust to using different quality thresholds (Figure S15). This graded measure is the pre-registered primary outcome measure for Experiment NR that recruited inactive misinformation sharers (ITT b = 0.011 [-0.011, 0.032], p = .320; ATT b = 0.017 [-0.018, 0.052], p = .330).

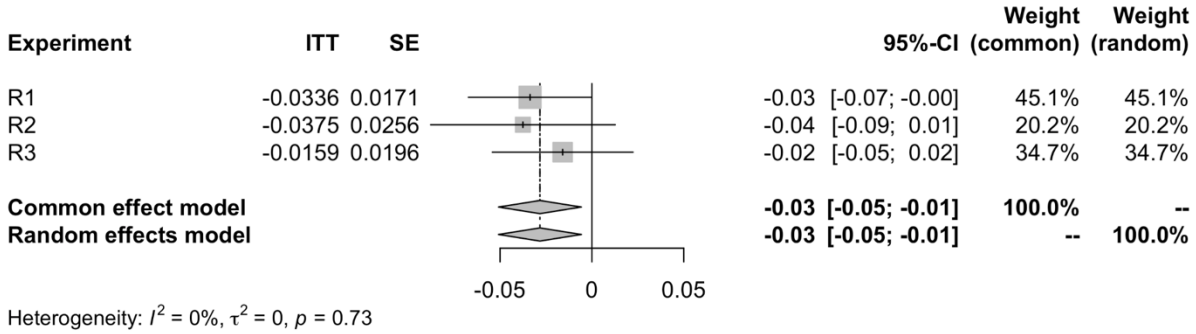

Figure S13*. Treatment effect (ITT) on reduction in misinformation sharing and meta-analytic effect. Experiments R1 and R3 used a graded approach to scoring domain quality. Experiment R2 counted the number of relevant low quality hashtags shared.*

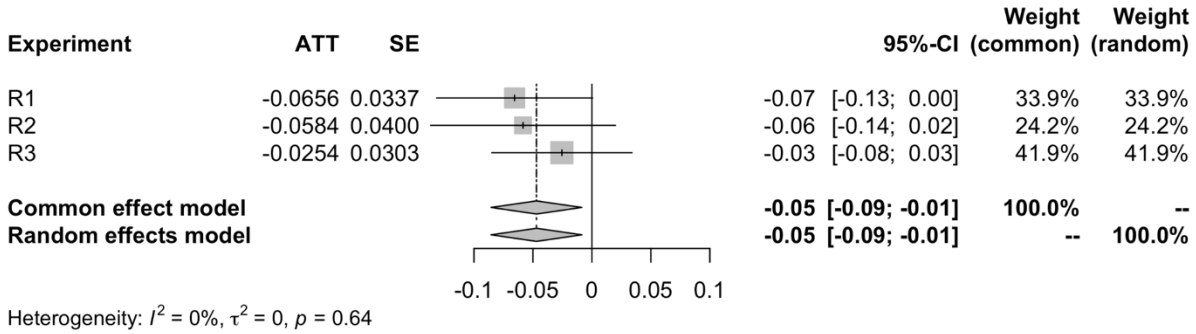

Figure S14*. Treatment effect (ATT) on reduction in misinformation sharing and meta-analytic effect. Experiments R1 and R3 used a graded approach to scoring domain quality. Experiment R2 counted the number of relevant low quality hashtags shared.*

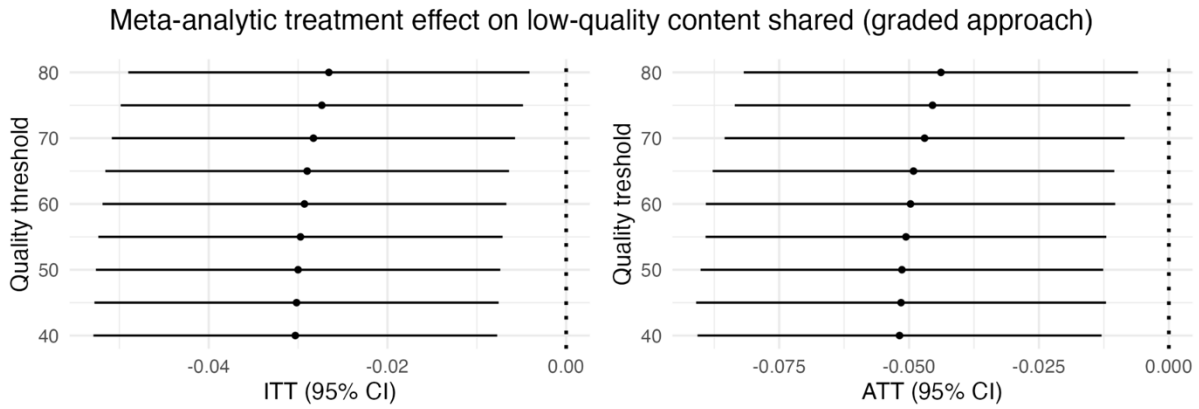

Figure S15*. Meta-analytic treatment effect for different domain quality thresholds (graded approach).*

S2.2.6 Fraction Low-Quality Domains Shared

We also examine whether the treatment reduced the *proportion* of low-quality content shared by users—defined as the number of tweets low-quality domains shared divided by the total number of retweets shared by each user. This analysis was conceptualized as a way to test whether the treatment was specifically affecting sharing of low quality content, versus content more generally. We later decided that simply comparing the treatment effect on count of low quality retweets to the treatment effect on count of all other retweets (as reported in the main text) was a better/more direct measure of this quantity. However, because this fraction analysis was pre-registered, we report it here.

There was no significant effect of the treatment on this outcome across experiments R1, R2, and R3 (ITT b = -0.0002 [-0.0006, 0.0002], p = .385; ATT b = -0.0003 [-0.001, 0.0004], p = .454; see Figures S16, S17, and S18) or in experiment NR (ITT b = -0.0004 [-0.001, 0.0001], p = .095; ATT b = -0.0007 [-0.002, 0.0001], p = .097).

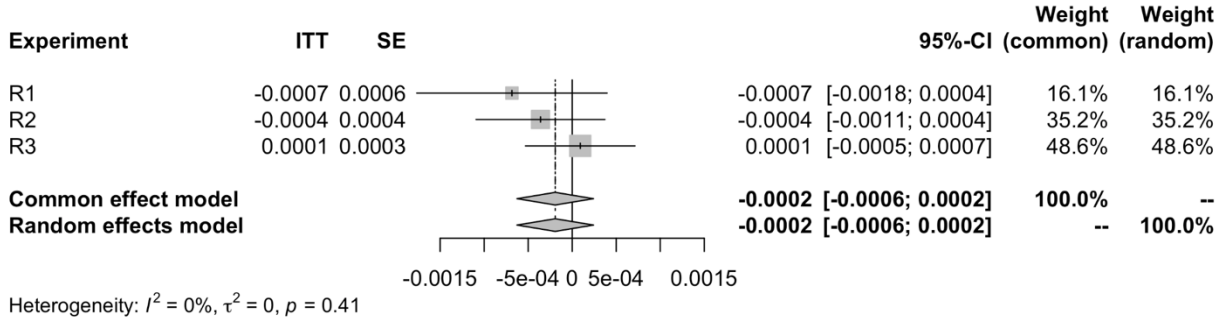

Figure S16*. Treatment effect (ITT) on reduction in misinformation sharing and meta-analytic effect. The outcome was fraction low-quality domains shared: (number of tweets with low-quality domains shared / all retweets). Estimates for each study were obtained from fixed-effects OLS.*

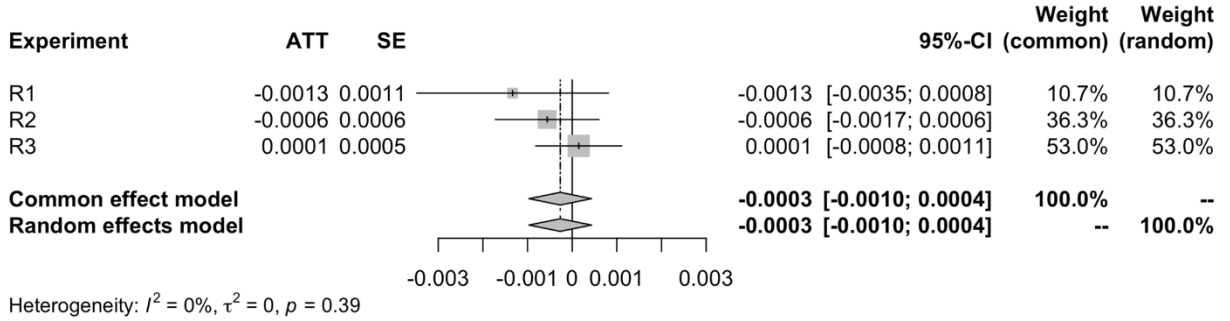

Figure S17*. Treatment effect (ATT) on reduction in misinformation sharing and meta-analytic effect. The outcome was fraction low-quality domains shared: (number of tweets with low-quality domains shared / all retweets). Estimates for each study were obtained from fixed-effects OLS.*

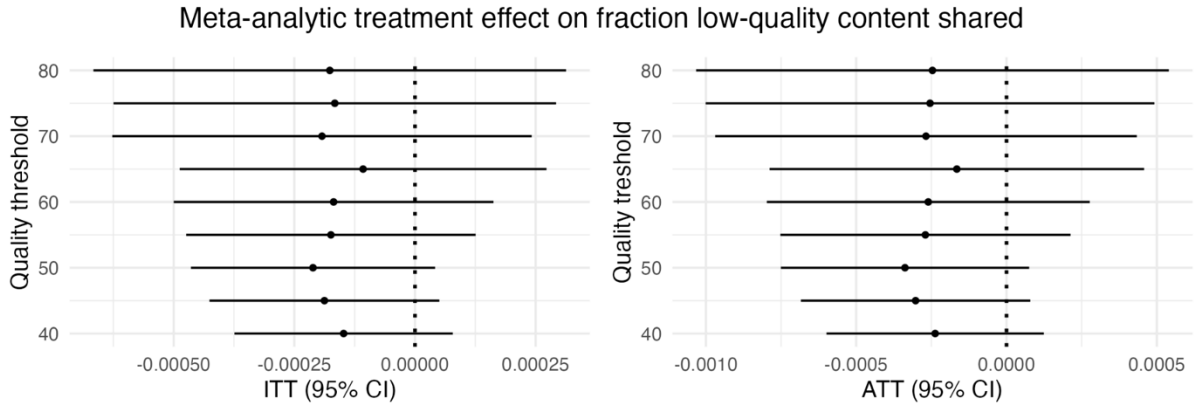

Figure S18*. Meta-analytic treatment effect for different domain quality thresholds. The outcome was fraction low-quality domains shared: (number of tweets with low-quality domains shared / all retweets).*

S2.2.7 Winsorization

In the main text, we winsorized variables by replacing values above the 95th percentile of all values with the 95th percentile of all values. Winsorization is helpful and often necessary because social media sharing data tend to have heavy right skews, and importantly, outliers (i.e., the few users who share large amounts of content) that can bias the causal effect estimates (but quasi-Poisson models can help to partly address these issues). When we winsorize at the 99th percentile, we still find that users in the treatment shared less low quality content than users in the control (ITT b = -0.036 [-0.065, -0.007], p = .014; ATT b = -0.062 [-0.112, -0.012], p = .016; see Figures S19 and S20). The results are relatively robust to using different quality thresholds (Figure S21).

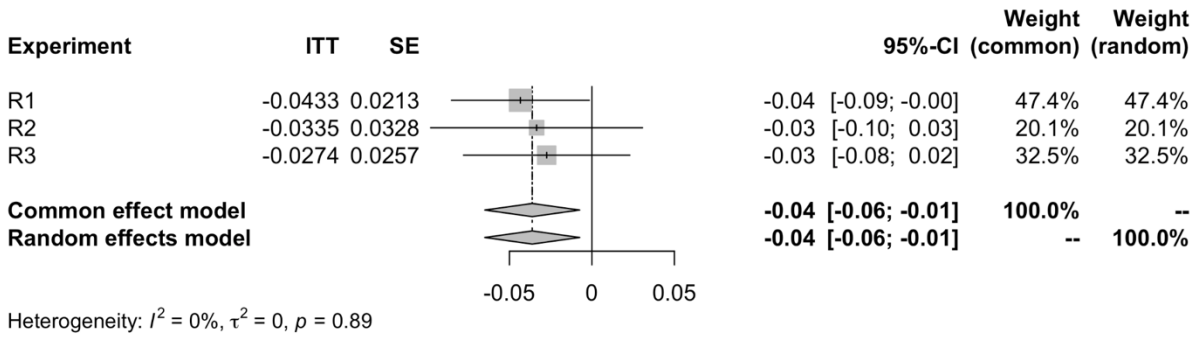

Figure S19*. Treatment effect (ITT) on reduction in misinformation sharing and meta-analytic effect. Experiments R1 and R3 used a graded approach to scoring domain quality. The outcome and pre-treatment variables are winsorized at the 99th percentile.*

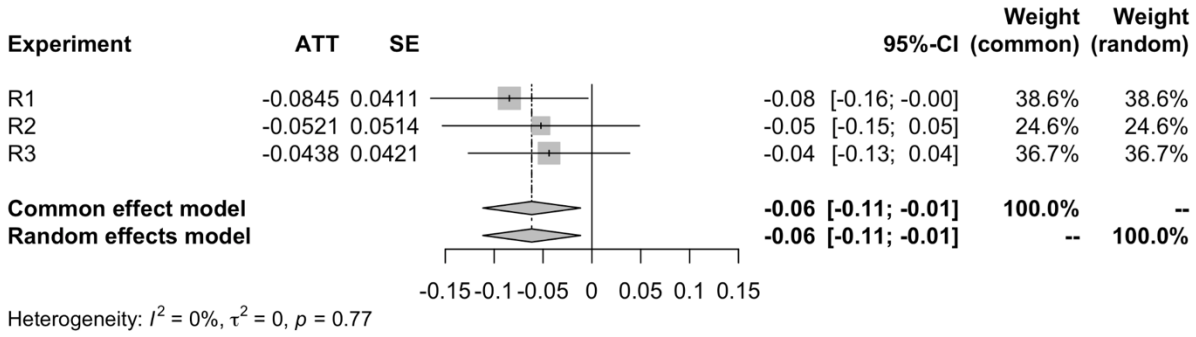

Figure S20*. Treatment effect (ATT) on reduction in misinformation sharing and meta-analytic effect. Experiments R1 and R3 used a graded approach to scoring domain quality. The outcome and pre-treatment variables are winsorized at the 99th percentile.*

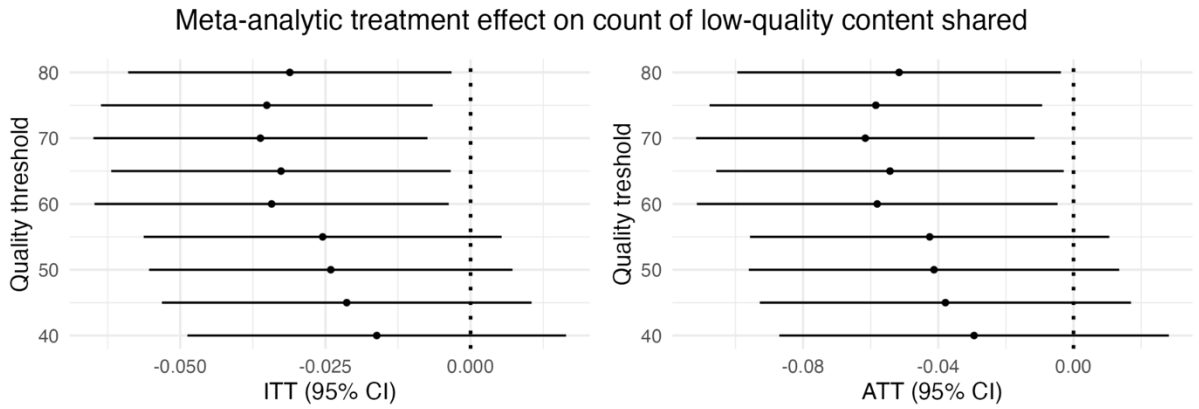

Figure S21. *Meta-analytic treatment effect for different domain quality thresholds. The outcome and pre-treatment variables are winsorized at the 99th percentile.*

Similarly, when we do not winsorize, we find that users in the treatment shared less low quality content than users in the control (ITT b = -0.04 [-0.08, -0.002], p = .038; ATT b = -0.07 [-0.14, -0.004], p = .039; see Figures S22 and S23). The results are also relatively robust to using different quality thresholds (Figure S24).

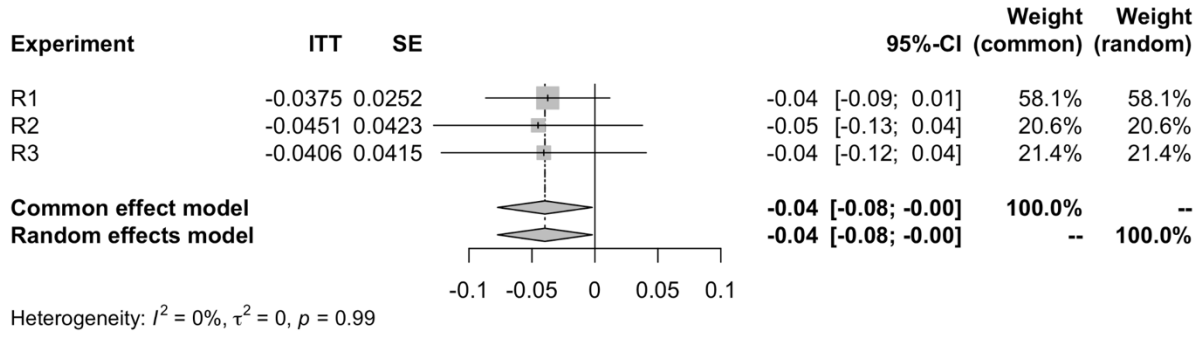

Figure S22. *Treatment effect (ITT) on reduction in misinformation sharing and meta-analytic effect. Experiments R1 and R3 used a graded approach to scoring domain quality. The outcome and pre-treatment variables are not winsorized.*

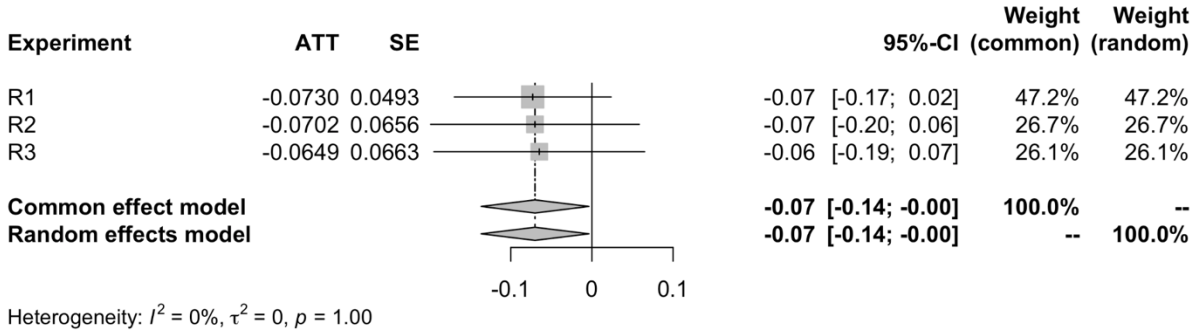

Figure S23. *Treatment effect (ATT) on reduction in misinformation sharing and meta-analytic effect. Experiments R1 and R3 used a graded approach to scoring domain quality. The outcome and pre-treatment variables are not winsorized.*

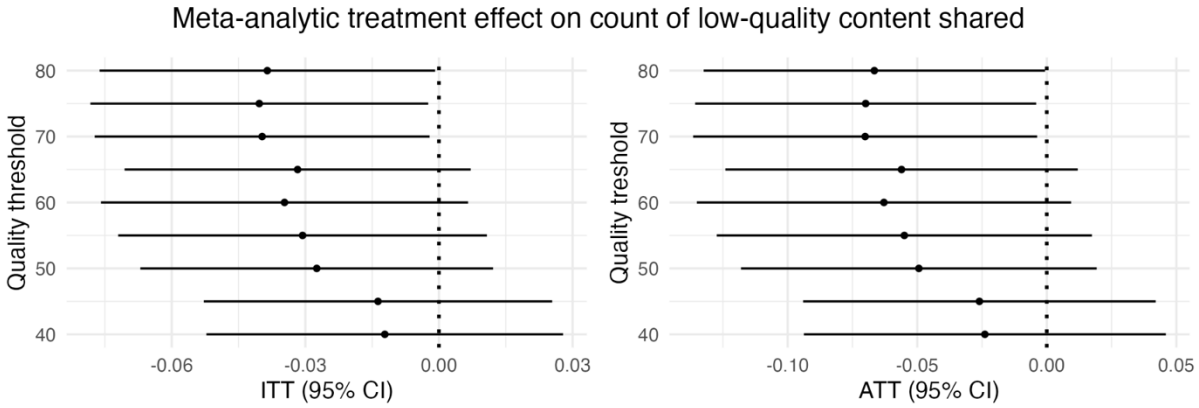

Figure S24*. Meta-analytic treatment effect for different domain quality thresholds. The outcome and pre-treatment variables are not winsorized.*

Supplementary References

S1. Ling, J. 5g and qanon: How conspiracy theorists steered canada's anti-vaccine trucker protest (internet archive). (2022).

S2. Marche, S. When the rage came for me (internet archive). *Atlantic* (2022).

S3. Taylor, S. Government communication on covid-19 contributed to 'freedom convoy' origin: Report (internet archive). *CTVNews* (2023).