Stanford | King Center on Global Development

Battling the Coronavirus Infodemic Among Social Media Users in Africa

Molly Offer-Westort

Leah R. Rosenzweig

Susan Athey

January, 2023

Working Paper No. wp2040

John A. and Cynthia Fry Gunn Building | kingcenter.stanford.edu

Battling the Coronavirus 'Infodemic' Among Social Media Users in Africa

Molly Offer-Westort[1], Leah R. Rosenzweig[2], and Susan Athey[3]

[1]Department of Political Science, University of Chicago; mollyow@uchicago.edu
[2]Development Innovation Lab, University of Chicago
[3]Stanford Graduate School of Business, Stanford University

January 12, 2023

Abstract:

During a global pandemic, how can we best prompt social media users to demonstrate discernment in sharing information online? We ran a contextual adaptive experiment on Facebook Messenger with users in Kenya and Nigeria and tested 40 combinations of interventions aimed at decreasing intentions to share misinformation while maintaining intentions to share factual posts related to COVID-19. We estimate precise null effects of showing users warning flags or suggesting related articles alongside misleading posts, tactics used by social media platforms. Instead, users share more discerningly when they are given tips for spotting misinformation or are nudged to consider information's accuracy, reducing misinformation sharing by 7.5% and 4.5% relative to control, respectively. We find significant heterogeneity in response to these treatments across users, indicating tips and the accuracy nudge affect outcomes through separate mechanisms. These low-cost, scalable interventions have the potential to improve the quality of information circulating online.

Amid the outbreak of the novel coronavirus (SARS-CoV-2), people around the world were also subjected to an “infodemic”—the spread of misinformation related to the virus. In the early days of the pandemic, people sought prevention techniques and remedies for the COVID-19 disease: In Nigeria, multiple people were hospitalized for chloroquine poisoning following statements by former President Trump suggesting the medication could be used to treat COVID-19 (Busari and Adebayo, 2020). In Iran, dozens of people died from alcohol poisoning after ingesting methanol supposedly due to the rumor that alcohol could prevent coronavirus (Haghdoost, 2020). While misinformation can be harmful along many dimensions, it is particularly dangerous in the context of a global pandemic.

This paper evaluates online interventions designed to deter users from sharing misinformation on social media without adversely affecting how they share true information on related topics. Our application focuses on information and misinformation about prevention and treatment for COVID-19, as we began the study in February 2021 before vaccines were widely available. Using targeted Facebook advertisements, we recruited a sample of social media users in Kenya and Nigeria, two of the three largest Facebook markets in sub-Saharan Africa (World Population Review, 2022). Users who clicked on our ads interacted with a Facebook Messenger chatbot to answer survey questions and receive randomized treatments. This mode of interaction kept social media users who selected into the study on the platform where they already engage with similar media posts.

Toward our goal of improving users’ sharing discernment, we implemented a multi-factorial adaptive experimental design to first learn the best-performing treatments. The adaptive design sequentially assigned treatment probabilities to privilege assignment to the most effective interventions and minimize assignment to ineffective or counterproductive interventions. Adaptive designs have two benefits: first, they improve outcomes for individuals during the experiment; second, in some circumstances they can lead to a more effective policy learned at the end of the experiment (Even-Dar et al., 2006; Caria et al., 2020; Kasy and Sautmann, 2021); policy learning with data from contextual adaptive experiments is especially challenging, however (Athey et al., 2022). Improving participants’ outcomes is important when conducting research on a sensitive topic like misinformation, where the literature has noted concerns of unintended negative consequences from well-intentioned interventions and debriefs (Swire-Thompson et al., 2020). In this paper, we demonstrate the use of an adaptive design to study effective countermeasures for the spread of misinformation while minimizing the likelihood of bad outcomes from assigning individuals to poor-performing treatments. Effective policy can be learned better because adaptive experiments allocate more individuals to the best treatments, improving how precisely those treatments are estimated; if there are many poorly performing treatments, this advantage is larger.

While traditional randomized experiments are more limited in the number of interventions they can test due to power constraints, our adaptive learning stage allowed us to sort through 40 unique treatment combinations. We examined two classes of interventions, randomized across users: (*headline-level*) treatments delivered on specific posts shown to treated users include flags or warning labels pinned on the article of interest; (*respondent-level*) messaging treatments delivered to treated users include tips for spotting fake news, training videos, and nudges. These two types of interventions vary in their cost and scalability. Whereas specific headline flags require fact-checking sources to keep up with the generation of new misinformation, delivering headline-agnostic interventions to users as they scroll social media sites is less costly and more easily scaled.

We conducted a second evaluation stage to estimate response under the most effective interventions from the learning stage, and to compare them against each other and to control. To do so, we recruited a new sample of Facebook users and randomly assigned them to: the pure control condition; the two most successful headline-level treatments; the two most successful respondent-level treatments; and a "learned targeted policy," a treatment program that assigned different respondent-level treatments to individuals based on their characteristics. The two headline-level treatments included in the evaluation phase were fact-check labels and accompanying posts with related articles (as Facebook has in the past provided under misleading or false posts, Ghosh 2017). The two respondent-level treatments included were Facebook's tips for spotting misinformation (Guess et al., 2020) and an accuracy nudge (Pennycook et al., 2021).

We find that the headline-level interventions do not perform better than control and estimate precise null effects for these treatments. Only a handful of scholars have examined Facebook's related articles policy (see also Bode and Vraga, 2015), hence more research is needed. Numerous experimental studies find fact checks to be effective at reducing users' *belief* in false stories (Nyhan and Reifler, 2010; Clayton et al., 2020; Brashier et al., 2021; Porter and Wood, 2021), but few focus on whether users share the information. Our null results on sharing lend further support to the notion that what users believe and what they share, though related, are distinct.

We do, however, see that individuals share more discerningly when given tips and an accuracy nudge. Other studies have found similar positive effects of accuracy prompts, including among quota-matched samples in 16 countries (Arechar et al., 2022) and in a meta-analysis of 20 accuracy experiments with a total sample size over 20,000 (Pennycook and Rand, 2022). Facebook tips have also been shown to effectively reduce belief in false headlines in the US and India (Guess et al., 2020), indicating that both treatments may be scalable solutions for the global misinformation challenge. Our results offer the further nuance that these interventions were also compared against numerous other treatments we tested.

We also evaluate whether there is heterogeneity across types of users (defined by pre-intervention characteristics) in benefit from treatments. Our learned targeted policy assigned users, based on their characteristics, to one of four respondent-level treatments to optimize an outcome measure that includes users' intentions to share both true and false posts.[1] In the evaluation stage, we find that this learned targeted policy improves over the control, but the benefit is imprecisely estimated, with effects arising from a reduction in false sharing offset by a smaller reduction in true sharing intentions. To focus on the most effective interventions, we further analyze the performance of an "alternative targeted policy" optimized to reduce the intention to share false posts, under which users are only assigned to the accuracy nudge or Facebook tips.

Expanding on existing studies and the debate on whether these two interventions operate through the same mechanism, we find that users with different characteristics have improved outcomes under their respective assignments in the alternative targeted policy. Specifically, a group of users with lower digital literacy and lower scientific knowledge do better under Facebook tips, whereas users with higher digital literacy and scientific knowledge have directionally improved responses under the accuracy nudge. The results suggest that there are benefits to targeting these interventions.

The findings have implications for fighting misinformation generally, and health misinformation specifically. We focused on what users share rather than what they believe, as exposure to misinformation can have harmful offline behavioral effects. Therefore, identifying effective interventions to deter users from sharing health falsities—and understanding which types of interventions are most effective for which types of users—may have life-saving effects. This study is not without limitations, however, particularly with respect to external validity, which we discuss in Section 2.

[1] The learned targeted policy is documented in an update to our pre-registration on the Open Science Framework registry: https://osf.io/ny2xc. We focus only on respondent-level treatments to yield a better comparison between the best overall fixed treatments (accuracy, Facebook tips) and a personalized targeted policy.

1 Results

Study sample We conducted this study with social media users in Kenya and Nigeria, two major English-language hubs of online communication in East and West Africa. We recruited social media users 18 years and older in these countries through targeted Facebook advertisements.[2] Users who clicked on our ads were prompted to start a conversation with our research page's Messenger chatbot. The chatbot serves both to collect survey responses and to deliver experimental interventions.

The study was conducted in two stages, each with unique participants: a learning stage, with 4,553 social media users, and an evaluation stage, with 10,683 social media users. In Supplementary Subsection S1.2 we report sample characteristics and compare with nationally representative Afrobarometer surveys.

Primary outcomes We operationalized sharing discernment using a combined response measure of sharing intentions. Both before and after treatment, participants in the experiment were shown a series of real social media posts about COVID-19 cures, treatments, and preventative best practices. More details on stimuli are provided in Methods Subsection 3.1. For each stimuli, users were separately asked whether they would share the stimuli through two channels: on Messenger and on their Facebook timeline. We control for pretest sharing (Davidian et al., 2005) and use an index of repeated measures (Broockman et al., 2017) to improve efficiency of effect estimation.

Our prespecified combined response measure is a weighted sum of times users said they would like to share true and misinformation stimuli over each channel. Our objective is to learn treatments that decrease sharing of false information without overly impeding sharing of true information; intentions to share false stimuli are given a weight of -1 and intentions to share true stimuli are given a weight of 0.5 in this measure.

We also report results for both types of stimuli separately: as the proportion of true and false stimuli participants planned to share across any channel, either Messenger or timeline. And we report sharing intentions disaggregated by sharing channel.

[2] See our advertisement in Supplementary Figure S1; for further details on targeted Facebook recruitment, see Rosenzweig et al. (2020).

Learning and evaluation stages We designed the learning stage to compare a large number of treatment conditions and to learn which of them were most effective on our prespecified combined response measure. We considered two classes of interventions: seven respondent-level interventions and four headline-level interventions. Supplementary Table S3 describes all of the interventions we tested. We used a multifactorial design where each class of intervention was treated as a separate multilevel factor, with a baseline control condition.

To assign treatment in the learning stage, we used a contextual adaptive assignment algorithm, a version of balanced linear Thompson sampling (Dimakopoulou et al., 2017, 2019), by which treatment assignment probabilities are updated based on covariates and the observed history of treatment and response. Under Thompson sampling, treatment is assigned under the Bayesian posterior probability that each treatment has the highest mean response. In linear Thompson sampling, this is generalized to allow the outcome to be a linear function of covariates.

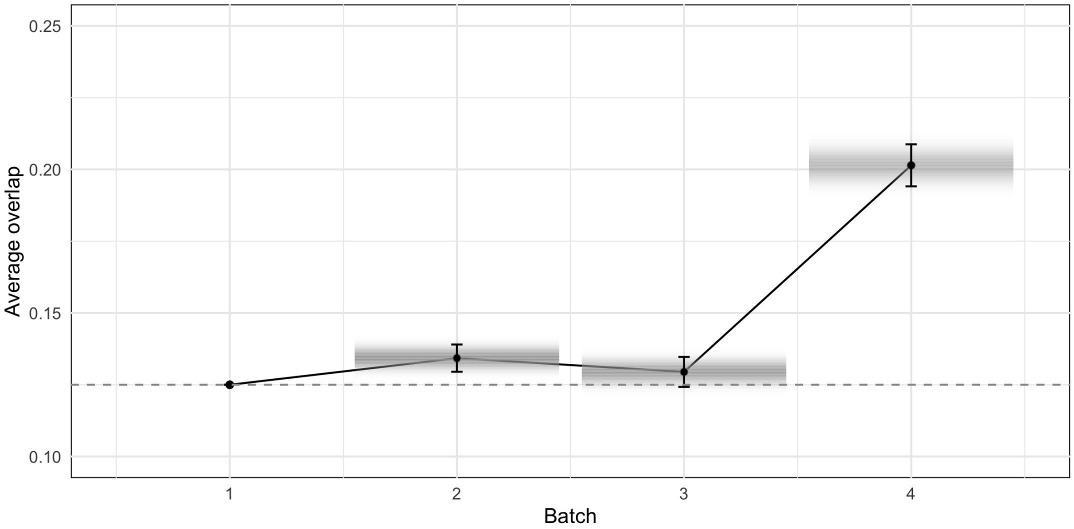


Figure 1. Evolution of on-policy probabilities during the adaptive experiment. The sample is users in the learning stage, total $n = 4,553$. The y-axis is the share of participants in the batch assigned to the respondent-level treatment assigned by the estimated targeted policy. The dashed grey line is the probability of being assigned to any respondent-level treatment under uniform random assignment.

This adaptive design allowed us to continue to learn which treatments were best, while reducing the probability that users were assigned to ineffective or harmful interventions, and increasing the probability that users were assigned to the most effective interventions. The inclusion of treatment-covariate interactions in the assignment model allows for the possibility that different interventions may be most effective for users with different covariate profiles. Figure 1 illustrates how the probability that users will be assigned to the alternative targeted policy increases over sequential batches of the adaptive experiment; as the experiment progresses, we assign users to the alternative targeted policy with higher probability. We include fixed probability floors to help us with policy learning (Zhan et al., 2021b).

In the learning phase, average response on our combined response function is -0.429 (SE = 0.020); we estimate average response under uniform random assignment would be -0.435 (SE = 0.027) using adaptively weighted estimators (Zhan et al., 2021a). While the difference is small and not statistically significant, the higher value under our algorithm indicates that the algorithm directionally improved cumulative regret over uniform assignment in the learning stage of the experiment. The relatively small improvement may be due to mis-assignment early in the adaptive experiment, as the algorithm may initially follow false leads before gathering more data and updating the response model.

From the data in the learning stage, we selected from the respondent-level and headline-level classes the two treatments associated with the highest estimated mean responses in each class separate from control, as measured on our combined response outcome. (Supplementary Figure S3 reports estimated response in the learning stage). These treatments were the accuracy nudge and Facebook tips (respondent-level) and fact-check and related articles (headline-level); examples of each are presented in Figure 5.

In the evaluation stage, we compared these most effective interventions to the control to obtain precise estimates of their effects. Treatment was assigned with equal probability to each of these or to the learned targeted policy, which assigned to each user the respondent-level treatment predicted to be best for them conditional on their covariate profile.

To learn a targeted policy, we fit a random forest model to the learning stage data, accounting for unequal assignment probabilities under adaptive assignment. Using this estimator we predict counterfactual responses on our combined response measure for each covariate profile in the evaluation stage. In the learned targeted policy, respondents may be assigned the accuracy nudge, Facebook tips, an emotion suppression prompt, or a video treatment. We also consider an alternative targeted policy, under which users can only be assigned to Facebook tips or the accuracy nudge. This alternative targeted policy is based on a causal forest model fit to the learning stage data, and is targeted only to minimize intention to share false posts. It is evaluated using data from evaluation stage users assigned to one of the two component treatments. There is 51% overlap between the learned targeted policy and the alternative targeted policy. The differences are largely due to a larger proportion of respondents assigned to Facebook tips instead of the accuracy nudge under the alternative policy; on average, the Facebook tips treatment is more effective at reducing false sharing, while the accuracy nudge performs better on the combined response measure. The alternative targeted policy was not preregistered. The following results are based on analysis of data from the evaluation stage.

Discernment under control Under the control condition, we see that users exhibit discernment in what types of stimuli they intend to share and over which channels they intend to share them. Overall, users report greater intentions to share true stimuli as compared to false stimuli on any channel (difference of 21.0 pp, SE = 0.9). Users are also more likely to want to share true stimuli publicly on their timeline as compared to by private message (difference of 3.2 pp, SE = 0.7), but it is the reverse for false stimuli (difference of -2.9 pp, SE = 0.6). The data suggest that even at baseline, participants are able to differentiate between true and false posts to some extent, and, when they do indicate wanting to share false posts, they make different decisions about how to share them as compared to true posts.

We also find that covariates are highly predictive of heterogeneity in baseline sharing behaviors. Understanding baseline heterogeneity in sharing behavior helps us to identify the greatest culprits of sharing misinformation. We target several key variables for examining heterogeneity, as well as for targeting interventions: age, gender, political allegiance, digital literacy, and scientific knowledge. We focus on these preregistered variables as they may already be measured by social media platforms (age, gender) or are of theoretical interest in social scientific research (political allegiance, digital literacy, and scientific knowledge).

Our data suggest that under control, younger subjects, those aligned with the ruling party, participants with low digital literacy, and those with low scientific knowledge have relatively lower outcomes on our combined response measure, indicating relatively higher false sharing intentions and/or lower true sharing intentions (see Table 1).

		False			True		
	Combined	Any sharing	Messenger	Timeline	Any sharing	Messenger	Timeline
Age							
Below median	−0.485	0.457	0.413	0.374	0.632	0.546	0.564
(n = 5,412)	(0.035)	(0.012)	(0.012)	(0.011)	(0.012)	(0.013)	(0.012)
Above median	−0.289	0.425	0.378	0.359	0.670	0.577	0.623
(n = 5,271)	(0.040)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Difference	−0.196***	0.031+	0.036*	0.015	−0.038*	−0.031+	−0.059***
	(0.053)	(0.017)	(0.017)	(0.016)	(0.017)	(0.018)	(0.017)
Gender							
Not male	−0.352	0.399	0.359	0.326	0.611	0.515	0.544
(n = 5,050)	(0.036)	(0.012)	(0.012)	(0.011)	(0.012)	(0.013)	(0.012)
Male	−0.420	0.479	0.428	0.403	0.687	0.603	0.638
(n = 5,633)	(0.038)	(0.012)	(0.012)	(0.012)	(0.011)	(0.012)	(0.012)
Difference	0.068	−0.079***	−0.069***	−0.077***	−0.076***	−0.088***	−0.094***
	(0.053)	(0.017)	(0.017)	(0.016)	(0.017)	(0.018)	(0.017)
Supports governing party							
Not aligned	−0.333	0.415	0.365	0.339	0.629	0.537	0.565
(n = 7,570)	(0.031)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
Aligned	−0.522	0.506	0.470	0.435	0.704	0.621	0.663
(n = 3,113)	(0.050)	(0.016)	(0.016)	(0.015)	(0.015)	(0.016)	(0.016)
Difference	0.189**	−0.091***	−0.104***	−0.096***	−0.074***	−0.084***	−0.099***
	(0.059)	(0.019)	(0.019)	(0.018)	(0.018)	(0.019)	(0.019)
Digital literacy index							
Below median	−0.531	0.494	0.449	0.419	0.674	0.587	0.621
(n = 5,443)	(0.038)	(0.012)	(0.012)	(0.012)	(0.011)	(0.012)	(0.012)
Above median	−0.240	0.387	0.340	0.312	0.627	0.534	0.564
(n = 5,240)	(0.037)	(0.012)	(0.012)	(0.011)	(0.012)	(0.013)	(0.013)
Difference	−0.291***	0.107***	0.108***	0.107***	0.048**	0.053**	0.057**
	(0.053)	(0.017)	(0.017)	(0.016)	(0.017)	(0.018)	(0.017)
Scientific knowledge index							
Below median	−0.442	0.458	0.413	0.383	0.658	0.564	0.597
(n = 5,677)	(0.036)	(0.012)	(0.012)	(0.011)	(0.012)	(0.012)	(0.012)
Above median	−0.327	0.423	0.376	0.349	0.643	0.558	0.589
(n = 5,006)	(0.039)	(0.013)	(0.012)	(0.012)	(0.012)	(0.013)	(0.013)
Difference	−0.115*	0.035*	0.037*	0.034*	0.016	0.005	0.009
	(0.053)	(0.017)	(0.017)	(0.016)	(0.017)	(0.018)	(0.017)

Table 1. **Heterogeneity in response under the control condition by selected covariates.** The sample is users in the evaluation stage, $n = 10,683$. Columns denote response measures, which include the combined response measure, a weighted sum of number of false sharing intentions (negatively weighted) and true sharing intentions (positively weighted); and for false and true posts separately, average propensity to share posts over any channel, over Messenger only, and on timeline only, reported in Subsection 3.1. Estimates are of mean response under the control condition and are produced from an augmented inverse probability weighted estimator, as described in Subsection 3.2, within specified subgroups. For contrasts only, under two-sided hypothesis tests: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; p-values are not adjusted for multiple testing.

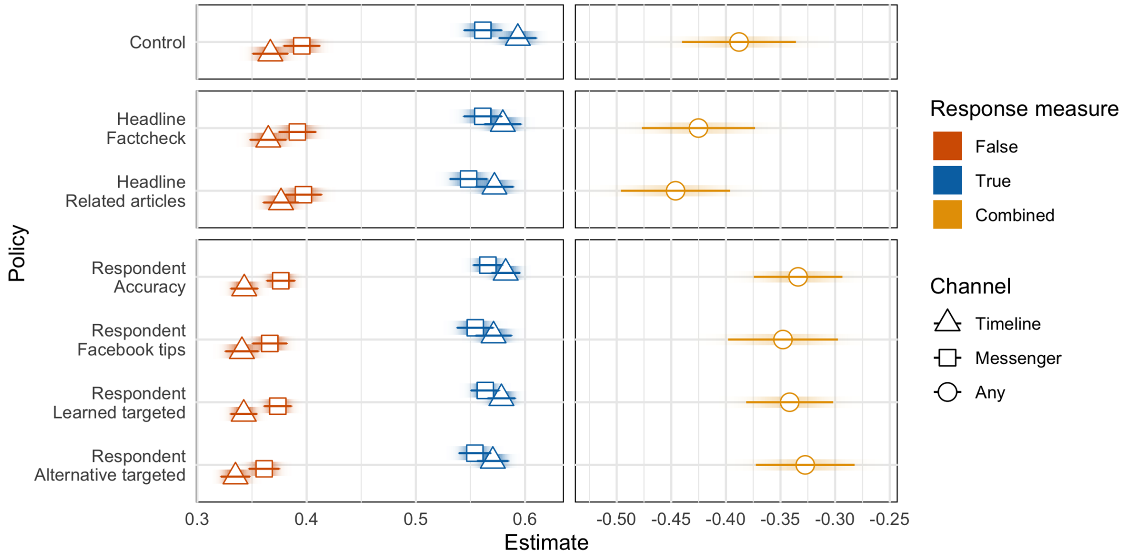


Figure 2. Response estimates. The sample is users in the evaluation stage, $n = 10,683$. Response measures are average propensity to share false and true posts over either channel, and a combined response measure, as reported in Subsection 3.1. Estimates are produced from an augmented inverse probability weighted estimator, as described in Subsection 3.2. Error bars represent 95% confidence intervals.

Main treatment effects Table 2 shows our prespecified comparisons of each evaluated treatment condition against the control. Our objective is to decrease intentions to share false information, while minimizing negative effects on intentions to share true information. To this end, an effective intervention would result in positive treatment effects for our combined response function, negative treatment effects on false sharing measures, and positive or neutral treatment effects on true sharing measures.

The two headline-level treatments are not effective at decreasing sharing of false stimuli while maintaining rates of sharing true stimuli. The related articles treatment directionally increases intention to share false stimuli as compared to control, although this estimate is not statistically distinguishable from zero at conventional significance levels. The fact-check treatment is associated with a decrease of 0.4 pp (SE = 1.2) in false sharing intentions as compared to control; the effect would need to be more than four times as large with the same degree of uncertainty for the confidence interval to exclude zero.

The respondent-level treatments, however, are effective. Facebook tips and the accuracy nudge increase the combined response measure by 0.040 (SE = 0.035) and 0.054 (SE = 0.032) relative to control, respectively. These effects are driven by decreases in false sharing of 3.4 pp (SE = 1.1) for Facebook tips and 2.0 pp (SE = 1.0) for the accuracy nudge. Effects on true sharing are not distinguishable from zero at conventional significance levels for either treatment. These interventions speak to the debate on whether misinformation spreads because people are not paying attention or people do not have skills or information to spot it (Ecker et al., 2022).

		False			**True**		
	Combined	Any sharing	Messenger	Timeline	Any sharing	Messenger	Timeline
Headline treatment effects							
Factcheck	−0.037	−0.006	−0.004	−0.002	−0.003	0.000	−0.014
	(0.036)	(0.012)	(0.011)	(0.011)	(0.011)	(0.012)	(0.012)
Related articles	−0.058	0.006	0.002	0.010	−0.019	−0.013	−0.021
	(0.035)	(0.012)	(0.011)	(0.011)	(0.011)	(0.012)	(0.012)
Respondent treatment effects							
Accuracy	0.054*	−0.020*	−0.019*	−0.024**	−0.001	0.005	−0.011
	(0.032)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
Facebook tips	0.040	−0.034**	−0.029**	−0.026**	−0.015	−0.007	−0.022
	(0.035)	(0.011)	(0.011)	(0.011)	(0.011)	(0.012)	(0.011)
Learned targeted policy	0.050+	−0.022*	−0.022*	−0.026**	−0.006	0.002	−0.014
(maximizing combined response)	(0.031)	(0.010)	(0.010)	(0.009)	(0.010)	(0.010)	(0.010)
Alternative targeted policy	0.063*	−0.038***	−0.035***	−0.034***	−0.016	−0.006	−0.021
(minimizing any false sharing)	(0.033)	(0.011)	(0.010)	(0.010)	(0.011)	(0.011)	(0.011)
Control mean	−0.388	0.441	0.396	0.367	0.651	0.561	0.593
	(0.027)	(0.009)	(0.008)	(0.008)	(0.008)	(0.009)	(0.009)

Table 2. Control response and treatment effect estimates. The sample is users in the evaluation stage, $n = 10,683$. Columns denote response measures, described in the note to Table 1 and in Subsection 3.1. The last row represents estimated mean response under the control condition; all other rows are estimated treatment effects in contrast with the control condition. Estimates are produced from an augmented inverse probability weighted estimator, as described in Subsection 3.2. For contrasts only, under one-sided hypothesis tests, as prespecified in preregistration: $^{+}\ p < 0.1$, $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$.

Heterogeneity in best policy While both Facebook tips and the accuracy nudge are effective, we also observe differences in how users respond to these two treatments. The learned targeted policy reported in Table 2 was fit prior to collecting evaluation data, and shows modest improvements over control. As false sharing intentions are more responsive to treatment as compared to true sharing intentions, however, we consider an alternative policy that targets propensity to share false information, and which assigns only the two most effective respondent-level treatments, Facebook tips and the accuracy nudge.

To learn this alternative targeted policy, we fit a causal forest model on the learning data, using the false sharing outcome measure. We use this model to predict counterfactual response under both Facebook tips and the accuracy for all individuals in the evaluation data, and calculate the difference between these two predicted responses. If the predicted difference indicates that the accuracy nudge more effectively reduces sharing of misinformation, our alternative targeted policy assigns the accuracy nudge; otherwise Facebook tips is assigned. When the alternative targeted policy is applied to the evaluation data, 59.6% of participants are assigned Facebook tips, which is the best uniform policy for decreasing sharing of false stimuli; the remaining respondents are assigned the accuracy nudge.

The causal forest model fit to the learning data is used only to determine assignment, and not for evaluation. We evaluate the alternative targeted policy on the evaluation data using the augmented inverse probability weighted estimator described in Subsection 3.2. Under this alternative targeted policy we achieve a treatment effect of -3.8 pp (SE = 1.1) in decreasing false sharing intentions (see Table 2). This is an improvement as compared to either Facebook tips or the accuracy nudge assigned uniformly (differences of -0.8 pp, SE = 0.8, and -1.8 pp, SE = 0.8, respectively).

The alternative targeted policy is also directionally more effective at moving the combined response function than the original learned targeted policy (difference of 0.013, SE = 0.019). When we restrict our targeted policy to only the most effective respondent-level treatments, and target only the outcome that is most responsive to treatment, we obtain a more effective policy; this is expected if the signal-to-noise ratio is unfavorable with a large number of candidate component treatments in the targeted policy.

Next, we present evidence on the benefits of targeting treatments, following Yadlowsky et al. (2021). Supposing hypothetically that a prespecified fraction of participants are to be allocated to accuracy rather than Facebook tips, we develop a targeted prioritization rule (following the same method for estimating counterfactual outcomes used to estimate the alternative targeted policy) for allocating subjects to the accuracy nudge. We compare expected outcomes under this prioritization rule to the case where the same fraction of participants are allocated to accuracy, but participants are selected randomly. Figure 3 illustrates this benefit to targeting, as we vary the percentage of participants allocated to accuracy. If we were limited to assigning the accuracy nudge to only 40 percent of the population and assigned Facebook tips to the remainder, false sharing intentions would be 3.9 pp lower (SE = 1.4) if we used the prioritization rule instead of random assignment. The overall rank-weighted average treatment effect, a weighted sum of the area under the curve in Figure 3, is -2.8 pp (SE = 1.3), using the targeting operator characteristic curve.

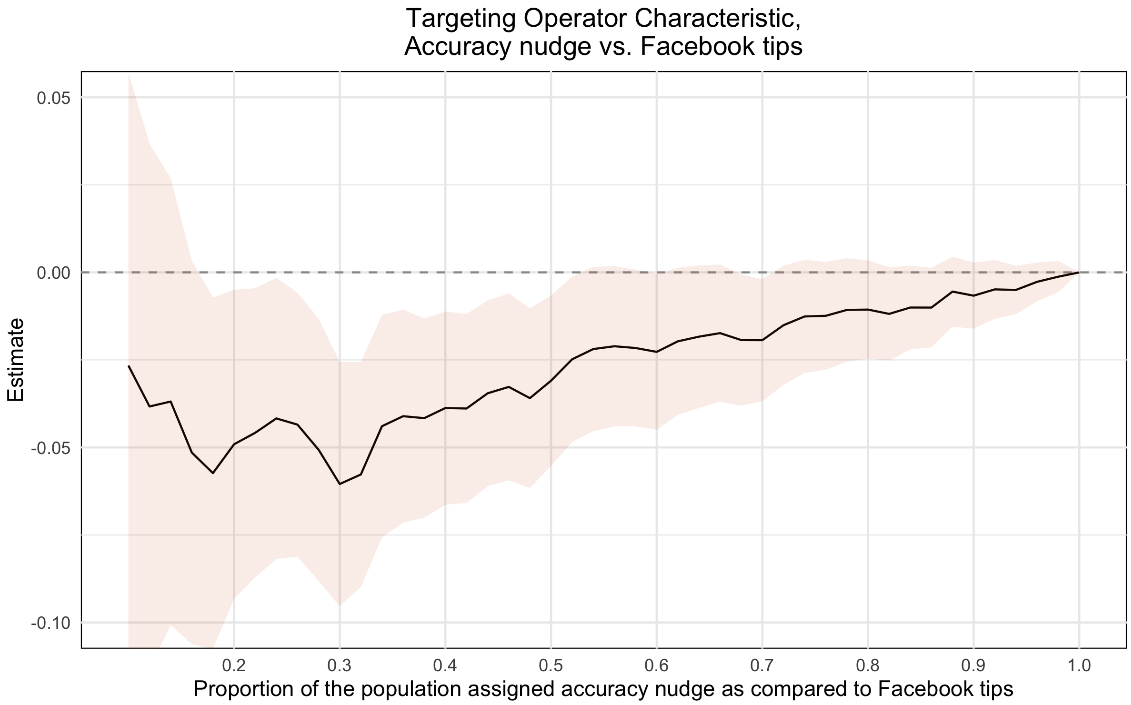


Figure 3. Targeting operator characteristic curve, comparing the accuracy nudge and Facebook tips. The policy is learned on the learning stage data. The sample for evaluation here is users in the evaluation stage, $n = 10,683$. The outcome measure is the difference in proportion of false stimuli participants reported wanting to share, either as a Facebook post or privately in Facebook Messenger, between the accuracy nudge and Facebook tips. The *y*-axis represents differences in this measure if the users receiving the accuracy nudge were assigned according to a prioritization rule, as compared to at random. The shaded region shows the 95% confidence interval.

In Table 3, we see that the alternative targeted policy has appropriately assigned participants to the respective respondent-level conditions: on average participants assigned to receive Facebook tips under the policy intend to share false information at lower rates under Facebook tips as compared to the accuracy nudge (difference of 3.5 pp, SE = 1.4); the reverse is true directionally for participants assigned the accuracy nudge under the policy (difference of -1.7 pp, SE = 1.6). However, the accuracy nudge does not perform significantly better in this subgroup than Facebook tips, and we cannot reject the null hypothesis that there is no difference between the alternative targeted policy and Facebook tips. Because the alternative targeted policy is optimized to minimize false sharing, we see smaller relative differences in true sharing and the combined measure.

		False			**True**		
	Combined	Any sharing	Messenger	Timeline	Any sharing	Messenger	Timeline
Optimal assignment == Accuracy nudge (n = 4,312)							
Accuracy	−0.044	0.339	0.295	0.276	0.638	0.534	0.574
	(0.033)	(0.010)	(0.010)	(0.009)	(0.011)	(0.011)	(0.011)
Facebook Tips	−0.088	0.356	0.308	0.294	0.639	0.540	0.577
	(0.037)	(0.012)	(0.012)	(0.011)	(0.014)	(0.013)	(0.013)
Difference	0.044	−0.017	−0.012	−0.018	0.000	−0.006	−0.004
	(0.050)	(0.016)	(0.015)	(0.015)	(0.018)	(0.017)	(0.017)
Optimal assignment == Facebook tips (n = 6,371)							
Accuracy	−0.530	0.476	0.431	0.388	0.658	0.588	0.588
	(0.026)	(0.009)	(0.009)	(0.008)	(0.009)	(0.009)	(0.009)
Facebook Tips	−0.524	0.442	0.406	0.373	0.634	0.564	0.567
	(0.035)	(0.011)	(0.011)	(0.010)	(0.011)	(0.011)	(0.011)
Difference	−0.007	0.035*	0.025+	0.016	0.024+	0.024+	0.021
	(0.043)	(0.014)	(0.014)	(0.013)	(0.014)	(0.014)	(0.014)

Table 3. Response under counterfactual uniform respondent treatment conditions, by alternative targeted policy assignment. The sample is users in the evaluation stage, $n = 10,683$. Estimates are of mean response under the two respondent-level treatments. Columns denote response measures, described in the note to Table 1 and in Subsection 3.1. Estimates are produced from an augmented inverse probability weighted estimator, as described in Subsection 3.2, within specified subgroups. For contrasts only, under two-sided hypothesis tests: $^{+}\ p < 0.1$, $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$.

Previous research has raised the question of whether Facebook tips and accuracy nudges both operate along the mechanism of increasing attention to accuracy, as suggested for the accuracy nudge by Pennycook et al. (2019), or rather whether Facebook tips improve ability to evaluate stimuli, as proposed by Guess et al. (2020). The heterogeneity we find in treatment effects between the two groups (difference of 5.2 pp, SE = 2.1) suggests, however, that there are differences in how users respond to these treatments and that different types of people respond differently to each treatment.

To better understand differences in the types of people that are most responsive to each of these interventions, we compare differences across our selected covariates. The 40.4% of participants assigned to the accuracy nudge under the alternative targeted policy are, on average, more digitally literate and more likely to have more scientific knowledge; they are also less likely to support the governing party, and more likely to be male (see Figure 4).

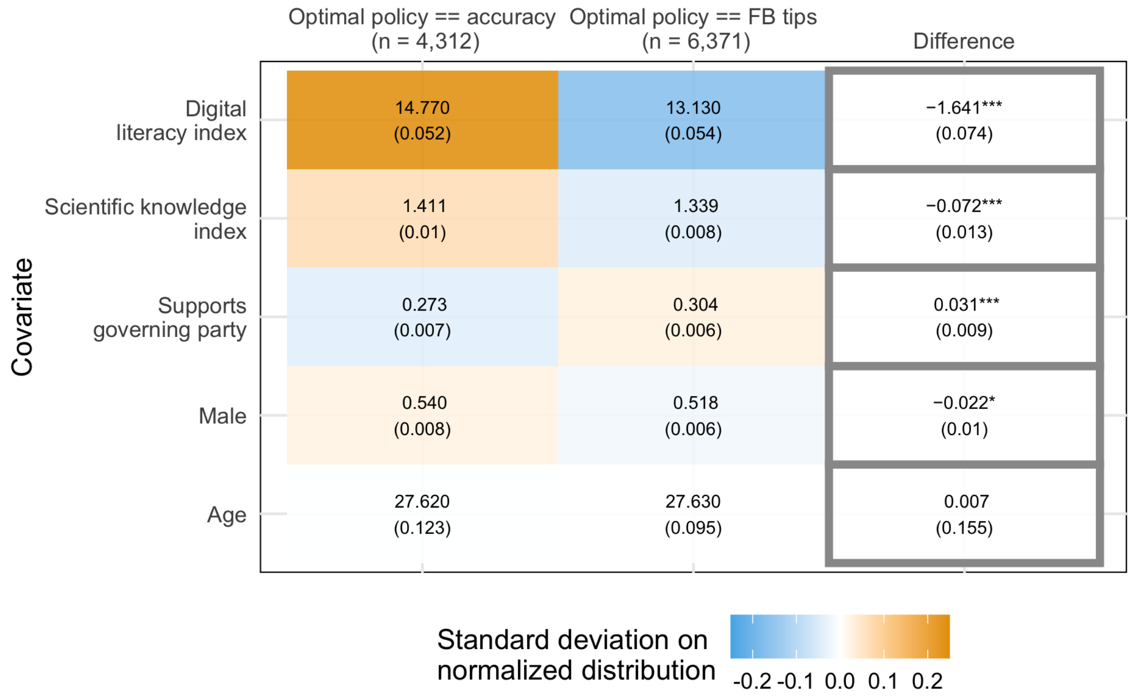


Figure 4. Selected covariate means for participants assigned to the accuracy nudge as compared to participants assigned to Facebook tips under the alternative targeted policy. The sample is users in the evaluation stage, $n = 10,683$. Covariates are ordered by size of standardized deviation between the two groups.

Sharing channel Overall, the Facebook tips treatment has directionally larger effects on mitigating false sharing intentions relative to the accuracy nudge (difference of −1.3 pp, SE = 1.0), but it also directionally reduces true sharing intentions (difference of −1.4 pp, SE = 1.0); this combination of effects results in the accuracy nudge scoring better on our combined response measure (difference of −0.014, SE = 0.03). To further investigate variations in how these two treatments operate, we consider the secondary dimension of our response measurement: sharing channel.

Facebook tips and the accuracy nudge are both effective at moving false sharing intentions on the timeline (−2.6 pp, SE = 1.1; −2.4 pp, SE = 1.0, respectively) and on Messenger

(-2.9 pp, SE = 1.1; -1.9 pp, SE = 1.0, respectively) relative to control. The Facebook tips treatment is directionally relatively more effective at reducing false sharing intentions on Messenger (difference of -1.0 pp, SE = 1.0).

This difference in effects by channel may speak to the mechanisms by which these two treatments work. We may suppose that one reason people share false posts is that they are not able to discern between true and false information. As noted, we do see evidence of discernment under control, where participants share false stimuli less on both channels relative to true stimuli. If treatments help users learn how to discern false stimuli from true stimuli, as is the objective of the Facebook tips, we should see effects both on timeline and on Messenger for false sharing intentions. We would also predict that these effects would be relatively larger for participants who are less able to differentiate false from true stimuli under control, as we discuss in the next section.

However, if sharing of false stimuli were merely due to users misattributing truth to some proportion of false stimuli, all else equal, under control we should expect that sharing rates by channel for false stimuli would be proportional to those for true stimuli. Rather, we see variation in users' preferred channel for sharing between true and false stimuli under control. An alternative mechanism through which treatment affects outcomes might be that the treatments are shifting attention to the accuracy of stimuli, as has been proposed for the accuracy nudge. For users who are already able to discern between true and false stimuli, it is ambiguous how this shift in attention should inform relative effects on the channel by which participants share stimuli. It may be that users are wary of sharing false posts publicly on their timeline out of fear for reputational costs they would incur from peers if they were caught sharing misinformation (Altay et al., 2022), but they may still believe that the posts are of interest or value to share with individual contacts. If increased attention to accuracy highlights concerns about reputational costs of publicly sharing false stimuli, this would result in larger relative effects on timeline as compared to Messenger false sharing intentions, as we see under the accuracy nudge.

We include additional analyses for heterogeneous treatment effects by individual covariates in Supplementary Subsection S2.2. The covariates and tests reported there are a subset of those pre-specified in our design document. We report tests of all pre-registered hypotheses alongside our online preregistration documents.

2 Discussion

This study provides evidence from two of the largest Facebook populations in sub-Saharan Africa that online interventions delivered via a Facebook Messenger chatbot are effective at improving sharing discernment. We show that out of numerous interventions tested targeting both specific posts and users generally, both Facebook tips and an accuracy nudge improve sharing discernment, largely by reducing intentions to share false posts. We find that headline-level treatments are ineffective. This study brings comparative data to the global problem of health misinformation, which to date draws primarily on empirical evidence from samples in the US, Canada, and Europe.

This study, like others of its kind, has limitations. First, our goal was to identify interventions that are effective among the population of social media users in Kenya and Nigeria. We were limited, however, in our recruitment methods to engaging with those who clicked on our Facebook advertisements to participate in the study. Recruiting actual social media users on the platform has advantages in validity relative to convenience samples, laboratory experiments, and opt-in survey panels. We cannot say, however, how users who decided to participate in our study differ on unobservables from the general population in these countries.

Second, interacting with participants of a study and delivering interventions in the course of a survey experiment cannot perfectly capture how users would react to real interventions delivered on the platform. Although still artificial, our approach of delivering the survey and interventions through a Facebook Messenger chatbot provides greater realism than interventions delivered on survey platforms like Qualtrics. The nature of our survey experiment means that participants were aware they were part of a study (rather than an on-platform field experiment, for example, where consent may be waived by IRB or implicitly provided when users agree to the terms and conditions). Therefore, it is possible that participants' responses could be driven by experimenter demand effects.

To address experimenter demand effects, we embedded treatments in a longer survey block about general social media usage. If users' post-treatment responses were, however, based on perceptions of what researchers want, we might expect high digital literacy users to be the most savvy to survey objectives and treatment effects to be largest for this group. Instead, we observe the reverse. The variation in treatment effects between Messenger and Timeline also provides evidence against experimenter demand effects: if users were only responding to perceived experimenter objectives, we might expect effects to be uniform across channels.

Finally, misinformation studies that focus on sharing behavior as the main outcome are constrained by ethical considerations of contributing to the ecosystem of misinformation by allowing participants to *actually share* false posts. This study, like most others, instead uses measures of sharing *intentions*. While scholars have found that intentions are correlated with online sharing behavior (Mosleh et al., 2020), measuring intentions rather than real behavior remains a limitation of scholarship in this area. In this study, we directly asked participants "Do you want to share this post on your timeline/on Messenger?" rather than posing a hypothetical question, although we did not immediately give users the opportunity to share posts. When we debriefed participants at the end of the study, we told them which posts were false and explained that was why they could not share those posts. We gave participants an opportunity to share true posts they had said they wanted to share.

Acknowledging these limitations, we believe this study offers insights useful for fighting online misinformation globally. The key insight is that low-cost and scalable accuracy nudges and tips for spotting misinformation delivered to users as they scroll social media can be effective in diverse contexts. This study provides evidence that such interventions are more effective than many others tested by researchers and used by platforms. Platforms may be more likely to deliver such interventions knowing that they help reduce sharing of misinformation without harming sharing of true information (perhaps one proxy for user engagement). Policymakers and platforms may also consider targeting interventions to those prone to sharing misinformation; they can avoid wasting resources or risking a worse user experience by *not* directing such interventions to groups for whom they are ineffective.

3 Methods

3.1 Data and recruitment

Our sample is recruited from Facebook users in Kenya and Nigeria, two of Facebook's top three largest user bases in sub-Saharan Africa (ITNews, 2016), with a combined user base of 30–35 million users ages 18 years and older.[3] We used targeted Facebook advertisements to improve balance on age and gender. After users clicked on our ads offering airtime for taking a survey (see Figure S1), they started a conversation with our page's Messenger

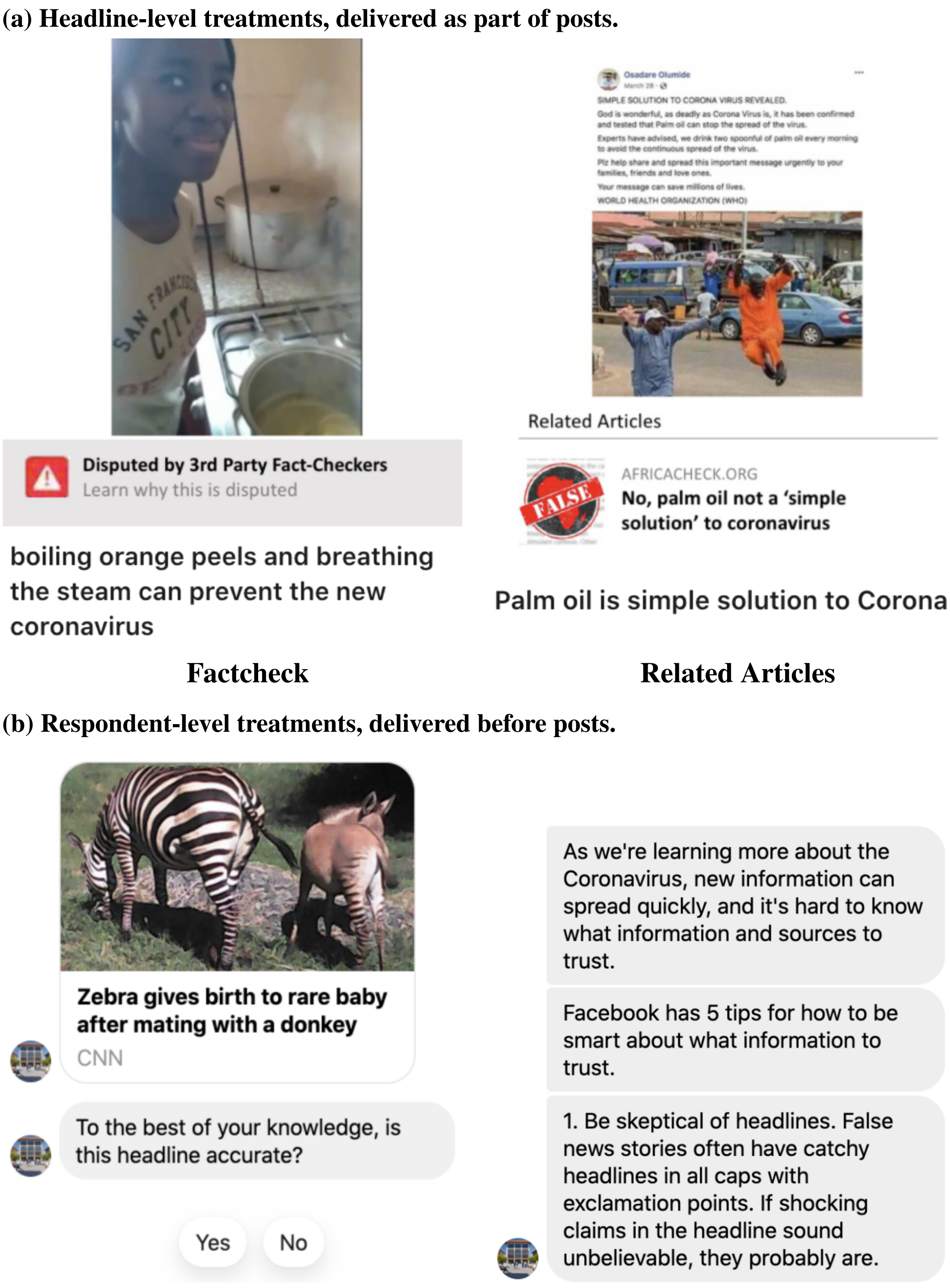
[3] Reported on the audience insights tool on Facebook's advertising platform.

chatbot. Participants who completed the survey received compensation in the form of mobile phone airtime (equivalent to about $0.50) sent to their phone.

Stimuli Each participant saw four posttreatment stimuli, two true and two false in a random order. For each stimuli, we asked participants two questions: if they wanted to share it (privately) in Facebook Messenger and if they wanted to share it (publicly) on their timeline. The stimuli include true information, sourced from the WHO, the Nigeria Center for Disease Control, the National Emergency Response Committee in Kenya, and the Ministry of Health in both countries. The false posts were sourced from AFP, Poynter, and AfricaCheck websites' lists of misinformation that had appeared online; the misinformation was fact-checked in Kenya and Nigeria since the start of the pandemic.

Treatments We considered two types of treatments, both randomized at the user-level: headline-level interventions applied to stimuli, and respondent-level interventions, including behavioral nudges and trainings targeted to the participants themselves. In the evaluation stage, we tested two of each type of intervention against control, along with a learned targeted policy composed of four of the respondent-level treatments.

The selected uniform treatments were the accuracy nudge and Facebook tips (respondent-level) and fact checks and related articles (headline-level). The accuracy nudge asked participants to tell us whether they thought a separate post, unrelated to COVID, was accurate or not (Pennycook et al., 2020). The Facebook tips treatment provided participants with ten tips Facebook has for how to be smart about what information to trust. These tips include being skeptical of headlines, watching for unusual formatting, checking the evidence, and looking at other reports, among others. The full text of the Facebook tips is presented in Supplementary Subsection S1.4. The fact-check treatment included a warning label on false stimuli, modeled on one used by Facebook for its third-party fact-checking program. The related articles treatment was also modeled on a program tested by Facebook, which paired disputed articles with articles on the same topic from validated sources. Examples of each are presented in Figure 5.



(a) Headline-level treatments, delivered as part of posts.

Disputed by 3rd Party Fact-Checkers

Learn why this is disputed

boiling orange peels and breathing the steam can prevent the new coronavirus

Factcheck

Osadare Olumide

March 28 ·

SIMPLE SOLUTION TO CORONA VIRUS REVEALED.

God is wonderful, as deadly as Corona Virus is, it has been confirmed and tested that Palm oil can stop the spread of the virus.

Experts have advised, we drink two spoonful of palm oil every morning to avoid the continuous spread of the virus.

Plz help share and spread this important message urgently to your families, friends and love ones.

Your message can save millions of lives.

WORLD HEALTH ORGANIZATION (WHO)

Related Articles

AFRICACHECK.ORG

No, palm oil not a 'simple solution' to coronavirus

Palm oil is simple solution to Corona

Related Articles

(b) Respondent-level treatments, delivered before posts.

Zebra gives birth to rare baby after mating with a donkey

CNN

To the best of your knowledge, is this headline accurate?

Accuracy

As we're learning more about the Coronavirus, new information can spread quickly, and it's hard to know what information and sources to trust.

Facebook has 5 tips for how to be smart about what information to trust.

1. Be skeptical of headlines. False news stories often have catchy headlines in all caps with exclamation points. If shocking claims in the headline sound unbelievable, they probably are.

Facebook Tips

Figure 5. Headline- and respondent-level treatments tested in the evaluation phase.

3.2 Empirical strategy

For both the learning and the evaluation stages of our study, we conduct estimation both accounting for unequal treatment assignment probabilities and adjusting for covariates.

To estimate average response under counterfactual treatment conditions and average treatment effects, we use a generalized augmented inverse probability weighted estimator (Robins et al., 1994). Scores are calculated separately for the learning and evaluation data. The learning data is used for estimating targeted treatment assignment policies, but we evaluate these learned targeted policies separately on the evaluation data.

The scores for the augmented inverse probability weighted estimator are calculated as

$$\Gamma_i^{AIPW}(w) := \hat{\mu}_i(X_i;w) + \frac{\mathbf{1}\{W_i = w\}}{e_i(X_i;w)} \left(Y_i - \hat{\mu}_i(X_i;w)\right), \tag{1}$$

where $\hat{\mu}_i(X_i;w)$ is a conditional means model, conditional on covariates X_i and categorical treatments $W_i \in \mathbf{W}$. Observed response for individual i is represented by Y_i. Treatment assignment probabilities are represented by $e_i(w) := \Pr[W_i = w \mid X_i = x]$. We estimate the conditional means model using a random forest as implemented by the `grf` page in R statistical software (Tibshirani et al., 2020).

The estimator is a substitution estimator, so we are able to predict counterfactual response for units under each of the different treatment conditions, and estimate average response under each condition by taking the averages of respective scores across the relevant units. To account for nonnormality of the estimator on the learning data, we use adaptive weights, described in Zhan et al. (2021a). Consequently for the learning data, the AIPW scores are weighted using evaluation weights, $h_i(w)$,

$$Q_i^h(w) := \frac{\frac{1}{N}\sum_{i=1}^N h_i(w)\Gamma_i(w)}{\sum_{i=1}^N h_i(w)}. \tag{2}$$

We use the contextual stabilized variance weights described by Zhan et al. (2021a). For the evaluation data, we aggregate scores to estimate $\mathrm{E}[Y_i(w)]$ as

$$Q_i^{AIPW}(w) := \frac{1}{N}\sum_{i=1}^{N} \Gamma_i^{AIPW}(w). \tag{3}$$

Contrasts are estimated by taking differences in (weighted) scores; estimation of standard errors follows the implementation in Tibshirani et al. (2020). Covariates used for adjustment are described in further detail in Table S4.

4 Acknowledgements

We received advertising credits for this study from Facebook Health and funding from the Golub Capital Social Impact Lab and Office of Naval Research grant N00014-19-1-246. For exceptional research assistance, we thank Zelin (James) Li, Ricardo Ruiz, Undral Byambadalai, and Haotian (Tony) Zong. We thank Laura Jakli, Shelby Grossman, Tanu Kumar, Alex Siegel, and Justine Davis for feedback and comments, as well as the participants of the seminar series of the Development Innovation Lab at the Becker Friedman Institute. We thank James Kiselik for editorial assistance.

References

Altay, S., Hacquin, A.-S., and Mercier, H. (2022). Why do so few people share fake news? It hurts their reputation. *New Media & Society*, 24(6):1303–1324.

Arechar, A. A., Allen, J. N. L., Cole, R., Epstein, Z., Garimella, K., Gully, A., Lu, J. G., Ross, R. M., Stagnaro, M., Zhang, J., et al. (2022). Understanding and reducing online misinformation across 16 countries on six continents.

Athey, S., Byambadalai, U., Hadad, V., Krishnamurthy, S. K., Leung, W., and Williams, J. J. (2022). Contextual bandits in a survey experiment on charitable giving: Within-experiment outcomes versus policy learning. *arXiv preprint arXiv:2211.12004*.

Bago, B., Rand, D. G., and Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General*.

Bago, B., Rosenzweig, L. R., Berinsky, A. J., and Rand, D. G. (2022). Emotion may predict susceptibility to fake news but emotion regulation does not seem to help. *Cognition and Emotion*, pages 1–15.

Bode, L. and Vraga, E. K. (2015). In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, 65(4):619–638.

Brashier, N. M., Pennycook, G., Berinsky, A. J., and Rand, D. G. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences*, 118(5):e2020043118.

Broockman, D. E., Kalla, J. L., and Sekhon, J. S. (2017). The design of field experiments with survey outcomes: A framework for selecting more efficient, robust, and ethical designs. *Political Analysis*, 25(4):435–464.

Busari, S. and Adebayo, B. (2020). Nigeria records chloroquine poisoning after Trump endorses it for coronavirus treatment. *CNN, Facts First*.

Caria, S., Kasy, M., Quinn, S., Shami, S., Teytelboym, A., et al. (2020). An adaptive targeted field experiment: Job search assistance for refugees in Jordan. *CESifo Working Paper*.

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., et al. (2020). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, 42(4):1073–1095.

Costa, M., Schaffner, B. F., and Prevost, A. (2018). Walking the walk? Experiments on the effect of pledging to vote on youth turnout. *PLOS One*, 13(5):e0197066.

Cotterill, S., John, P., and Richardson, L. (2013). The impact of a pledge request and the promise of publicity: A randomized controlled trial of charitable donations. *Social Science Quarterly*, 94(1):200–216.

Davidian, M., Tsiatis, A. A., and Leon, S. (2005). Semiparametric estimation of treatment effect in a pretest–posttest study with missing data. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 20(3):261.

Dimakopoulou, M., Athey, S., and Imbens, G. (2017). Estimation considerations in contextual bandits. *arXiv preprint arXiv:1711.07077*.

Dimakopoulou, M., Zhou, Z., Athey, S., and Imbens, G. (2019). Balanced linear contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3445–3453.

Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., and Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29.

Even-Dar, E., Mannor, S., Mansour, Y., and Mahadevan, S. (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(6).

Ghosh, S. (2017). Facebook will show people anti-fake news articles when they post false stories. *Insider.com*. url: https://www.insider.com/facebook-related-articles-feature-will-show-you-anti-fake-news-2017-8.

Gilens, M. (2001). Political ignorance and collective policy preferences. *American Political Science Review*, pages 379–396.

Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2(3):271–299.

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., and Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the united states and india. *Proceedings of the National Academy of Sciences*, 117(27):15536–15545.

Haghdoost, Y. (2020). Alcohol poisoning kills 100 Iranians seeking virus protection. *Bloomberg Markets*.

ITNews, A. (2016). Top 10 African countries with the most Facebook users. *ITNews Africa*. url: https://www.howwe.ug/news/lifestyle/14791/top-10-countries-with-the-most-facebook-users-in-africa.

Kasy, M. and Sautmann, A. (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica*, 89(1):113–132.

Martel, C., Pennycook, G., and Rand, D. G. (2019). Reliance on emotion promotes belief in fake news.

Mosleh, M., Pennycook, G., and Rand, D. G. (2020). Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter. *PLOS One*, 15(2):e0228882.

Nyhan, B. and Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., and Rand, D. G. (2019). Understanding and reducing the spread of misinformation online.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., and Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855):590–595.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., and Rand, D. G. (2020). Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, page 0956797620939054.

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

Porter, E. and Wood, T. J. (2021). The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proceedings of the National Academy of Sciences*, 118(37):e2104235118.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Rosenzweig, L. R., Bago, B., Berinsky, A. J., and Rand, D. G. (2021). Happiness and surprise are associated with worse truth discernment of COVID-19 headlines among social media users in Nigeria. *Harvard Kennedy School Misinformation Review*.

Rosenzweig, L. R., Bergquist, P., Hoffmann Pham, K., Rampazzo, F., and Mildenberger, M. (2020). Survey sampling in the global south using facebook advertisements.

Sanchez, C. and Dunning, D. (2021). Cognitive and emotional correlates of belief in political misinformation: Who endorses partisan misbeliefs? *Emotion*.

Swire-Thompson, B., DeGutis, J., and Lazer, D. (2020). Searching for the backfire effect: Measurement and design considerations. *Journal of Applied Research in Memory and Cognition*, 9(3):286–299.

Tibshirani, J., Athey, S., and Wager, S. (2020). *grf: Generalized Random Forests*. R package version 1.2.0.

World Population Review (2022). Facebook users by country 2022. url: https://worldpopulationreview.com/country-rankings/facebook-users-by-country.

Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., and Wager, S. (2021). Evaluating treatment prioritization rules via rank-weighted average treatment effects. *arXiv preprint arXiv:2111.07966*.

Zhan, R., Hadad, V., Hirshberg, D. A., and Athey, S. (2021a). Off-policy evaluation via adaptive weighting with data from contextual bandits. *arXiv preprint arXiv:2106.02029*.

Zhan, R., Ren, Z., Athey, S., and Zhou, Z. (2021b). Policy learning with adaptively collected data. *arXiv preprint arXiv:2105.02344*.

Supplementary Information

S1 Design and measurement SI.2

S1.1 Recruitment SI.2

S1.2 Sample Characteristics SI.3

S1.3 Survey instrument SI.4

S1.4 Treatments SI.4

S1.4.1 Facebook Tips SI.5

S1.5 Covariates SI.6

S1.6 Response measurement SI.9

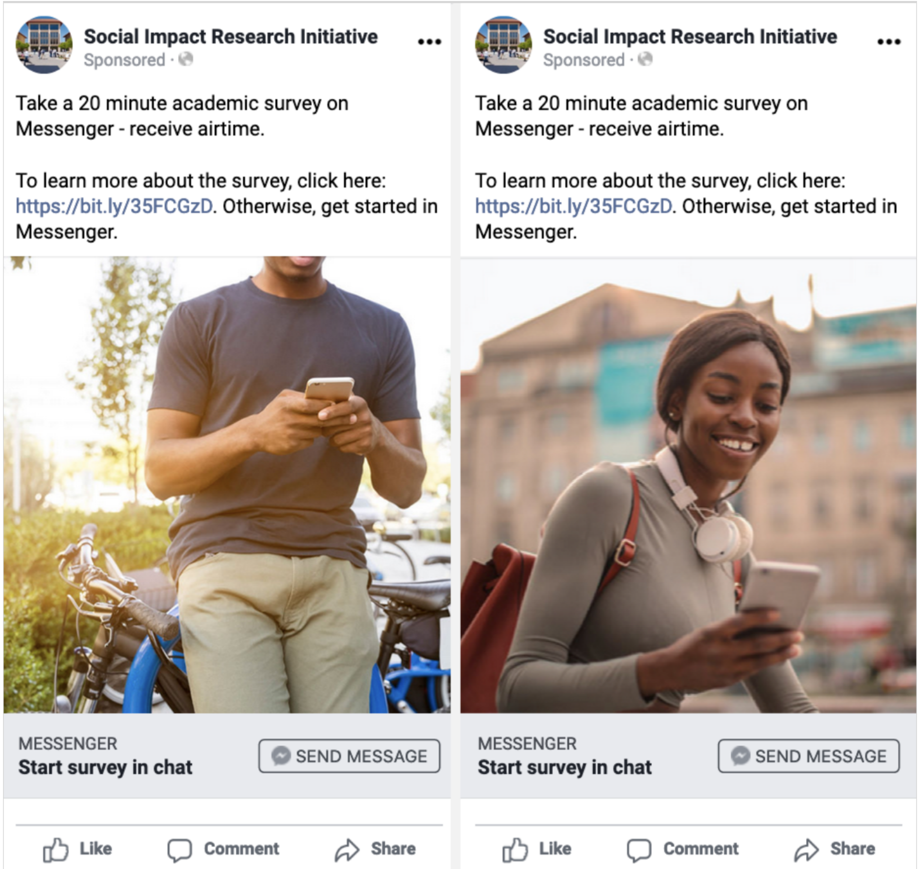
S2 Additional results SI.11

S2.1 Learning stage SI.11

S2.2 Evaluation stage SI.13

S1 Design and measurement

S1.1 Recruitment



Social Impact Research Initiative

Sponsored ·

Take a 20 minute academic survey on Messenger - receive airtime.

To learn more about the survey, click here: https://bit.ly/35FCGzD. Otherwise, get started in Messenger.

MESSENGER

Start survey in chat

Social Impact Research Initiative

Sponsored ·

Take a 20 minute academic survey on Messenger - receive airtime.

To learn more about the survey, click here: https://bit.ly/35FCGzD. Otherwise, get started in Messenger.

MESSENGER

Start survey in chat

Figure S1. Advertising image used for recruitment.

S1.2 Sample Characteristics

Table S1. Comparing Afrobarometer and Facebook samples from Kenya

	Afrobarometer Mean	Afrobarometer SE	Facebook Sample Mean	Facebook SE	Difference
Age	36.16	0.34	29.38	0.09	-6.79
Has cash income	0.48	0.01	0.38	0.01	-0.10
Education level	4.85	0.05	7.17	0.02	2.32
Index of household possessions	2.93	0.04	3.83	0.02	0.89
Male	0.50	0.01	0.54	0.01	0.04
Supports governing party	0.30	0.01	0.12	0.00	-0.18
Christian	0.76	0.01	0.91	0.00	0.15
Muslim	0.09	0.01	0.04	0.00	-0.06
Urban	0.36	0.01	0.41	0.01	0.05

Note: The Facebook sample household asset index is re-coded to exclude a question about bike ownership, to match the household asset index in the Afrobarometer data. Analysis in the main paper includes this additional question in the index. Additionally, the Facebook sample governing party support variable is coded to only include affiliation with the governing party here, again to match the Afrobarometer data, which only asks about prospective voting. In the main analysis, the governing party support variable is coded as 1 if the respondent either responds that they feel affiliation with the governing party, or if they voted for a candidate from that party in the previous election.

Table S2. Comparing Afrobarometer and Facebook samples from Nigeria

	Afrobarometer Mean	Afrobarometer SE	Facebook Sample Mean	Facebook SE	Difference
Age	32.66	0.31	26.42	0.08	-6.24
Has cash income	0.49	0.01	0.39	0.01	-0.10
Education level	5.52	0.05	7.34	0.02	1.83
Index of household possessions	4.00	0.04	4.49	0.02	0.48
Male	0.50	0.01	0.50	0.01	0.00
Supports governing party	0.26	0.01	0.18	0.00	-0.08
Christian	0.55	0.01	0.68	0.01	0.13
Muslim	0.42	0.01	0.29	0.01	-0.14
Urban	0.44	0.01	0.60	0.01	0.16

Note: The Facebook sample household asset index is re-coded to exclude a question about bike ownership, to match the household asset index in the Afrobarometer data. Analysis in the main paper includes this additional question in the index. Additionally, the Facebook sample governing party support variable is coded to only include affiliation with the governing party here, again to match the Afrobarometer data, which only asks about prospective voting. In the main analysis, the governing party support variable is coded as 1 if the respondent either responds that they feel affiliation with the governing party, or if they voted for a candidate from that party in the previous election.

S1.3 Survey instrument

The survey script is available at this link:
http://bit.ly/facebook_survey_public

All of the stimuli (posts) used in the experiment are available at this link:
http://bit.ly/facebook_stimuli_public

S1.4 Treatments

Treatments 1, 2, 3, 8, 9, and 10 are derived from interventions currently being used by social media platforms including Facebook, Twitter, and WhatsApp. For instance, Guess et al. (2020) find that reading Facebook's tips for spotting untrustworthy news improved participants' ability to discern false from true headlines in the US and India. Treatment 11 (real information) is a similar headline-level treatment that *could* be adopted by industry partners. Rather than flags or warnings about misinformation, we test whether providing a simple true statement reduces sharing of false information. Existing research suggests that providing true information can sometimes influence individuals' attitudes and behaviors (Gilens, 2001). Treatments 4, 6, and 7 are taken from previous academic studies. Emotions (4) have been suspected to influence susceptibility to misinformation (Martel et al., 2019; Rosenzweig et al., 2021; Bago et al., 2022); our test evaluates one canonical method of emotion suppression as a way to reduce the influence of misinformation. The accuracy nudge treatment (6) was specifically found to be effective at reducing the sharing of COVID-19 misinformation among participants in the US. Our deliberation nudge treatment (7) was adapted from Bago et al. (2020) that found asking participants to deliberate was effective at improving discernment of online political information. The pledge treatment (5) was adapted from the types of treatments used by political campaigns to get subjects to pledge to vote or support a particular candidate (Costa et al., 2018). We vary whether the pledge is made in private (within the chatbot conversation) or in public (posted on the respondent's Facebook timeline) to test whether public pledges are more effective at influencing behavior than private ones Cotterill et al. (2013).

Shorthand Name	**Treatment Level**	**Treatment**
1. Facebook tips	Respondent	Facebook's "Tips to Spot False News"
2. AfricaCheck tips	Respondent	Africacheck.org's guide: "How to vet information during a pandemic"
3. Video training	Respondent	BBC video on spotting Coronavirus misinformation
4. Emotion suppression	Respondent	Prompt: "As you view and read the headlines, if you have any feelings, please try your best not to let those feelings show. Read all of the headlines carefully, but try to behave so that someone watching you would not know that you are feeling anything at all" (Gross, 1998).
5. Pledge	Respondent	Prompt: Respondents will be asked if they want to keep their family and friends safe from COVID-19, if they knew COVID-19 misinformation can be dangerous, and if they're willing to take a *public* pledge to help identify and call out COVID-19 misinformation online.
6. Accuracy nudge	Respondent	Placebo headline: "To the best of your knowledge, is this headline accurate?" (Pennycook et al., 2020, 2019).
7. Deliberation nudge	Respondent	Placebo headline: "In a few words, please say *why* you would or would not like to share this story on Facebook." [open text response]
8. Related articles	Headline	Facebook-style related stories: below story, show one other story that corrects a false news story
9. Factcheck	Headline	Indicates story is "Disputed by 3rd party fact-checkers"
10. More information	Headline	Provides a message and link to "Get the facts about COVID-19"
11. Real information	Headline	Provides a *true* statement: "According to the WHO, there is currently **no proven** cure for COVID-19."
12. Control	N/A	Control condition

Table S3. Full list of treatments run during the learning phase.

S1.4.1 Facebook Tips

The script for the Facebook tips respondent-level treatment is as follows:

As we're learning more about the Coronavirus, new information can spread quickly, and it's hard to know what information and sources to trust. Facebook has some tips for how to be smart about what information to trust.

1. Be skeptical of headlines. False news stories often have catchy headlines in all caps with exclamation points. If shocking claims in the headline sound unbelievable, they probably are.

2. Look closely at the link. A phony or look-alike link may be a warning sign of false news. Many false news sites mimic authentic news sources by making small changes to the link. You can go to the site to compare the link to established sources.

3. Investigate the source. Ensure that the story is written by a source that you trust with a reputation for accuracy. If the story comes from an unfamiliar organization, check their "About" section to learn more.

4. Watch for unusual formatting. Many false news sites have misspellings or awkward layouts. Read carefully if you see these signs.

5. Consider the photos. False news stories often contain manipulated images or videos. Sometimes the photo may be authentic, but taken out of context. You can search for the photo or image to verify where it came from.

6. Inspect the dates. False news stories may contain timelines that make no sense, or event dates that have been altered.

7. Check the evidence. Check the author's sources to confirm that they are accurate. Lack of evidence or reliance on unnamed experts may indicate a false news story.

8. Look at other reports. If no other news source is reporting the same story, it may indicate that the story is false. If the story is reported by multiple sources you trust, it's more likely to be true.

9. Is the story a joke? Sometimes false news stories can be hard to distinguish from humor or satire. Check whether the source is known for parody, and whether the story's details and tone suggest it may be just for fun.

10. Some stories are intentionally false. Think critically about the stories you read, and only share news that you know to be credible.

S1.5 Covariates

In all analyses, we include the pretest response strata for true and false stimuli and indicators for individual stimuli. For some continuous covariates that describe individual characteristics, such as education, we include an indicator flag if the respondent skipped the question; this is noted in the “Coded as” column. For others which require reflection or where there is a “correct” or “best” response, such as the Cognitive Reflection Test or the COVID-19 information measure, we code the index as 0 if the respondent chose not to answer any of the questions.

Covariate	**Response options**	**Coded as**
Gender	Male, Female, Nonbinary, Other	1 if male, 0 otherwise
Age	Integers	Continuous, flag if greater than 120
Education	No formal schooling, Informal schooling only, Some primary school, Primary school completed, Some secondary school, Secondary school completed, Post-secondary qualifications, Some university, University completed, Post-graduate	1:10, flag if missing
Geography	Urban, Rural	1 if urban, 0 otherwise
Religion	Christian, Muslim, Other/None	Indicators
Denomination (Christian)	Pentecostal, Other	Indicator (coded 1 if Pentecostal, 0 otherwise)
Religiosity (freq. of attendance)	Never, Less than once a month, One to three times per month, Once a week, More than once a week but less than daily, Daily	1:6, flag if missing
Locus of control	[See survey instrument for full list]	1:10, flag if missing
Index of scientific views	[See survey instrument for full questions and response options]	0:2, flag if missing
Digital Literacy Index	[Based on the first nine items of Guess et al. (2020)'s proposed measure, see survey instrument for full questions and response options]	0:24
Frequency of social media usage (x2)	[See survey instrument for full questions and response options]	0:3, flag if missing
Cognitive Reflection Test	[See survey instrument for full questions and response options]	0:3 (1 point for each correct response)
Index of household possessions	I/my household owns, Do not own [See survey instrument for items]	Continuous, sum of owned items, flag if all missing
Job with cash income	Yes, No	1 if yes
Number of people in household	Integers	Continuous, flag if missing
Political affiliation	Governing party v. opposition	Indicator (coded 1 if associate with or voted for candidate from governing party, 0 otherwise)
Concern regarding COVID-19	Not at all worried, Somewhat worried, Very worried	1:3, flag if missing
Perceived government efficacy on COVID-19	Very poorly, Somewhat poorly, Somewhat well, Very well	1:4, flag if missing
Strata of response to pretest stimuli	[Would share stimuli on timeline/via Messenger]	Indicators for strata (0:2) x (True + False = 2 types) × (timeline + Messenger = 2 channels)

Note: Regarding missingness flags, respondents must respond to chatbot questions to advance in the survey, but for contexts they may enter "skip" if they do not wish to answer a given question, with the exception of age, which we check is greater than 18.

Table S4. Covariates and response options

S1.6 Response measurement

We are primarily interested in decreasing sharing of harmful false information about COVID-19 cures and treatments, but we simultaneously wish to limit any negative impact on sharing of useful information about transmission and best practices from verified sources. In this case, we care more about the spread of false COVID cures because in an environment of fear and uncertainty, belief that a cure will work may not play a large role in whether an individual tries a particular treatment when no proven alternative exists. We measure sharing intentions with two questions asked after each post the user saw: 1) would you like to share this post on your timeline? 2) would you like to send this post to a friend on Messenger?

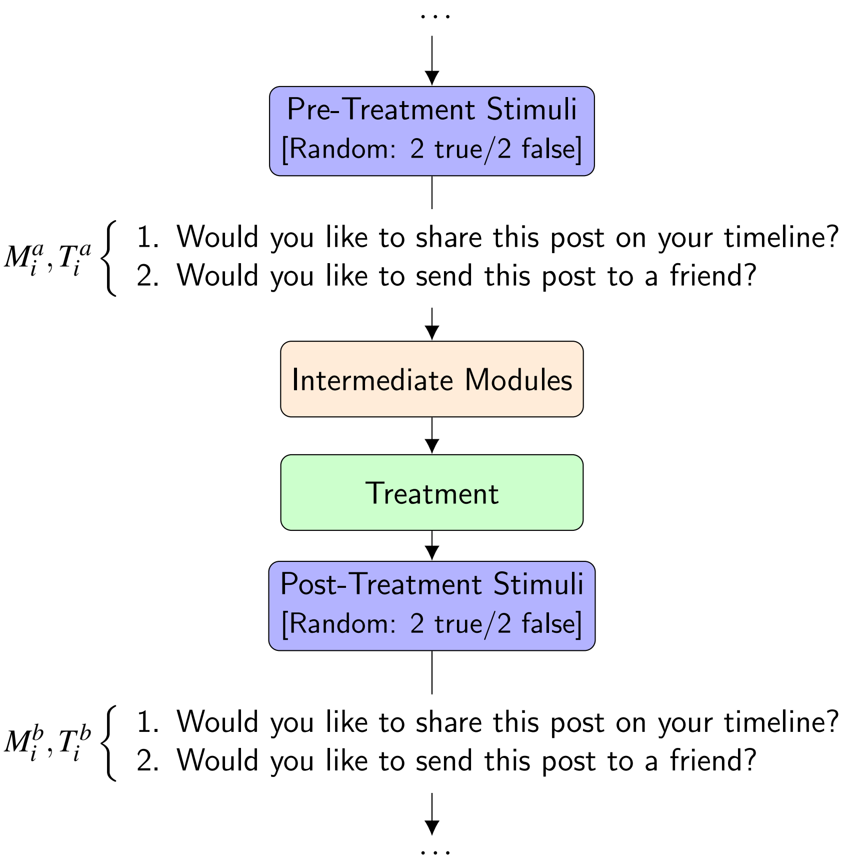

...
Pre-Treatment Stimuli
[Random: 2 true/2 false]
M_i^a, T_i^a
1. Would you like to share this post on your timeline?
2. Would you like to send this post to a friend?
Intermediate Modules
Treatment
Post-Treatment Stimuli
[Random: 2 true/2 false]
M_i^b, T_i^b
1. Would you like to share this post on your timeline?
2. Would you like to send this post to a friend?
...

Figure S2. Survey flow.

By using a pretest / posttest design (Davidian et al., 2005) as presented in Figure S2 and an index of repeated measures (Broockman et al., 2017), we aim to improve the efficiency of our effect estimation. Prior to treatment, we show participants four media posts from their country (two true and two false in random order) randomly sourced from our stimuli set. For each stimuli we ask the above self-reported sharing intention questions. Participants are then asked a series of questions about their media consumption and randomly assigned treatment according to the experimental design. If assigned to one of the respondent-level treatments, they are administered the relevant treatment. They are then shown four additional stimuli (two true and two false), selected from the remaining stimuli that they were *not* shown pretreatment. If the respondent is assigned a headline-level treatment, this treatment is applied only to the misinformation stimuli, as flags and fact-checking labels are not generally applied to true information from verified sources. For each of the stimuli we again ask the same self-reported sharing intention questions.

We code responses to the self-reported questions as one if the respondent affirms they want to share the post and zero otherwise. Let M_i^a be the sum of respondent i's pretest responses to the *misinformation* stimuli and let T_i^a be the sum of respondent i's pretest responses to the *true* informational stimuli. We denote the respective sums of post-treatment responses by M_i^b and T_i^b. By construction, $M_i^a, T_i^a, M_i^b, T_i^b \in \{0,1,2,3,4\}$.

We control for strata of pretest responses in our analyses. We formalize our response function in terms of posttest measures:

$$Y_i = -M_i^b + 0.5T_i^b.$$

This response function is the metric for which we optimize in our adaptive algorithm. Table S5 illustrates the values this combined response measure could take based on the number of intended true and false shares.

		True shares				
		0	1	2	3	4
	0	0.0	0.5	1.0	1.5	2.0
	1	−1.0	−0.5	0.0	0.5	1.0
False shares	2	−2.0	−1.5	−1.0	−0.5	0.0
	3	−3.0	−2.5	−2.0	−1.5	−1.0
	4	−4.0	−3.5	−3.0	−2.5	−2.0

Table S5. Combined response measure.

S2 Additional results

S2.1 Learning stage

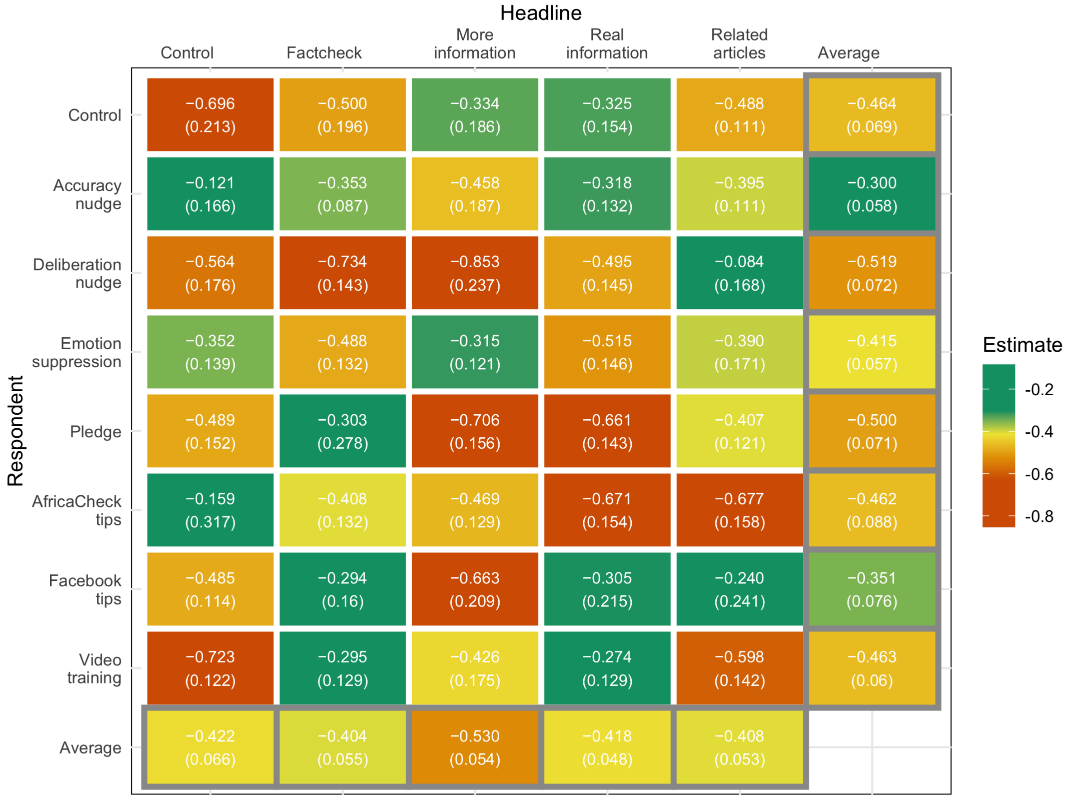


Figure S3. Learning stage estimates. The sample is users in the learning stage, total $n = 4,553$. Columns represent headline factors, rows respondent factors, and the intersections report treatment factor interactions and standard errors in terms of our combined response measure. Averages over rows and columns represent row- or column-wide averages, with interactions equally weighted. Estimates are produced from an augmented inverse probability weighted estimator, as described in Section 3.2.

The adaptive assignment privileges assignment to those interventions that are predicted to be most effective, down-weighting assignment to interventions that are predicted to perform poorly. This means that we collect more data about the interventions that are the most likely to succeed. It is important to note that adaptively collected data introduces additional challenges for policy learning (Zhan et al., 2021b); the exploitation of the bandit can eventually result in extreme probabilities of treatment assignment. However, this exploitation is an important ethical consideration in a setting where we are concerned about avoiding "backfire" from counter-productive interventions. The adaptive algorithm allows us to minimize these potentially harmful effects.

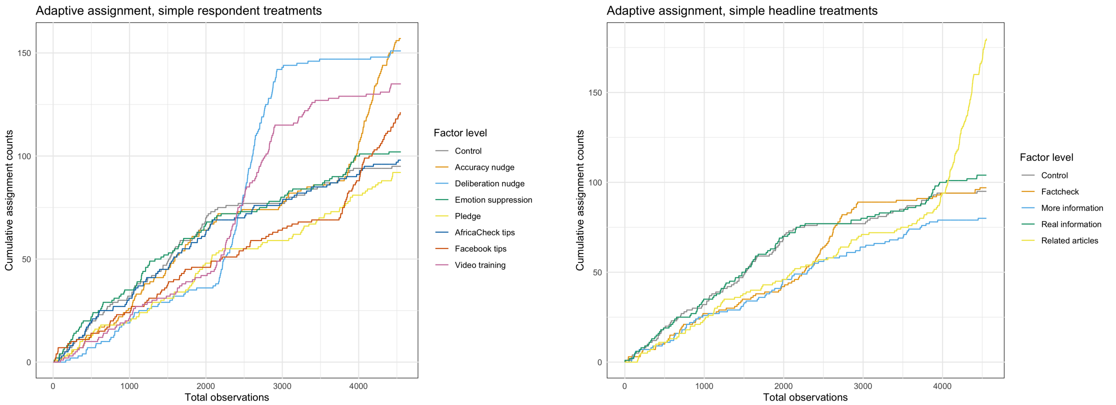


Figure S4. Cumulative treatment assignment during the learning phase for headline (left panel) and respondent (right panel) interventions. The sample is users in the learning stage, total $n = 4,553$. While the full design allows for all factor combinations, these plots illustrate assignment using the "pure" version of each factor, i.e., when the other factor is at the baseline control condition.

S2.2 Evaluation stage

Table 1 illustrates that users with low digital literacy, participants aligned with the ruling party, participants with low scientific knowledge, and younger participants intend to share more false stimuli under the control condition. For these "worst offenders," we find that assigning the respondent-level treatments on average decreases false sharing as compared to control among participants with low digital literacy (−4.0 pp, SE = 1.3), men (−2.7 pp, SE = 1.3), and participants with low scientific knowledge (−4.3 pp, SE = 1.3). (See Table S6.) The pooled respondent-level interventions do not reduce sharing of false posts among younger participants but do among older ones. Similarly, there is no effect of the pooled respondent treatments on false sharing among those aligned with the political party in power, but we do see a significant effect among those not aligned. However, *differences* in treatment effects across groups are for the most part only statistically significant when comparing users with low to those with high levels of scientific knowledge.

While only suggestive, these findings may reflect, as other studies have documented, that affective partisanship and motivated reasoning influence sharing of misinformation (Sanchez and Dunning, 2021). It is somewhat surprising, however, that even for less (blatantly) political information of COVID-19 best practices, these interventions are unable to move ruling party supporters.

The largest treatment effects on false sharing intentions were for users with below median digital literacy and below median scientific knowledge. For these users, like users on average, the Facebook tips treatment was more effective than the accuracy nudge (difference of 1.2 pp, SE = 1.3 for digital literacy; 1.6 pp, SE = 1.3 for scientific knowledge) (see Table S7 and Table S8).

For users with below median digital literacy and below median scientific knowledge, treatment effects under Facebook tips were driven by relatively larger effects on private sharing on Messenger as compared to public sharing on their timelines (difference of 1.2 pp, SE = 1.2 for digital literacy; 1.3 pp, SE = 1.2 for scientific knowledge), whereas for the accuracy nudge, effects on timeline as compared to messenger sharing are comparable for these groups. The overall larger effects on private Messenger sharing for Facebook tips as compared to the accuracy nudge are concentrated among these users. This may suggest that the Facebook tips treatment not only helps users to better differentiate between true and false stimuli, but for some types of users, it also makes them less likely to privately share stimuli that they already know is false.

		False			True		
	Combined	Any sharing	Messenger	Timeline	Any sharing	Messenger	Timeline
Age							
Below median	−0.001	−0.019	−0.020	−0.017	−0.025+	−0.017	−0.036**
(n = 5,412)	(0.039)	(0.013)	(0.013)	(0.012)	(0.014)	(0.014)	(0.014)
Above Median	0.097*	−0.035**	−0.029*	−0.033*	0.009	0.015	0.003
(n = 5,271)	(0.044)	(0.014)	(0.013)	(0.013)	(0.013)	(0.014)	(0.014)
Difference	0.098+	−0.016	−0.008	−0.017	0.034+	0.031	0.039*
	(0.059)	(0.019)	(0.018)	(0.018)	(0.019)	(0.019)	(0.019)
Gender							
Not male	0.032	−0.021	−0.021	−0.019	−0.024+	−0.007	−0.031*
(n = 5,050)	(0.041)	(0.014)	(0.013)	(0.012)	(0.014)	(0.014)	(0.014)
Male	0.061	−0.033*	−0.027*	−0.030*	0.006	0.004	−0.004
(n = 5,633)	(0.042)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
Difference	0.028	−0.013	−0.006	−0.011	0.030	0.011	0.027
	(0.059)	(0.019)	(0.018)	(0.018)	(0.019)	(0.019)	(0.019)
Supports governing party							
Not aligned	0.093**	−0.035**	−0.029**	−0.037***	−0.007	0.002	−0.015
(n = 7,570)	(0.035)	(0.011)	(0.011)	(0.011)	(0.011)	(0.012)	(0.012)
Aligned	−0.064	−0.008	−0.013	0.005	−0.011	−0.009	−0.020
(n = 3,113)	(0.055)	(0.017)	(0.017)	(0.017)	(0.017)	(0.018)	(0.017)
Difference	−0.156*	0.027	0.017	0.042*	−0.005	−0.011	−0.005
	(0.065)	(0.021)	(0.020)	(0.020)	(0.021)	(0.021)	(0.021)
Digital literacy index							
Below median	0.050	−0.044**	−0.040**	−0.034**	−0.024+	−0.013	−0.026*
(n = 5,443)	(0.042)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
Above median	0.044	−0.010	−0.008	−0.016	0.009	0.011	−0.007
(n = 5,240)	(0.042)	(0.013)	(0.013)	(0.012)	(0.014)	(0.014)	(0.014)
Difference	−0.006	0.033+	0.031+	0.018	0.033+	0.024	0.019
	(0.059)	(0.019)	(0.018)	(0.018)	(0.019)	(0.019)	(0.019)
Scientific knowledge index							
Below median	0.084*	−0.045***	−0.047***	−0.038**	−0.026*	−0.014	−0.033*
(n = 5,677)	(0.040)	(0.013)	(0.013)	(0.012)	(0.013)	(0.014)	(0.013)
Above median	0.005	−0.007	0.001	−0.010	0.012	0.013	0.002
(n = 5,006)	(0.044)	(0.014)	(0.013)	(0.013)	(0.014)	(0.014)	(0.014)
Difference	−0.080	0.038*	0.048**	0.028	0.039*	0.028	0.034+
	(0.059)	(0.019)	(0.018)	(0.018)	(0.019)	(0.019)	(0.019)

Table S6. Heterogeneity in treatment effects under averaged respondent-level treatments by selected covariates. The sample is users in the evaluation stage, $n = 10,683$. Columns denote response measures, which include the combined response measure, a weighted sum of number of false sharing intentions (negatively weighted) and true sharing intentions (positively weighted); and for false and true posts separately, average propensity to share posts over any channel, over Messenger only, and on timeline only, as reported in Subsection 3.1. Estimates are of treatment effects averaged across the two respondent-level treatments, in contrast with the control condition. Estimates are produced from an augmented inverse probability weighted estimator, as described in Subsection 3.2, within specified subgroups. Under two-sided hypothesis tests: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

		False			True		
	Combined	Any sharing	Messenger	Timeline	Any sharing	Messenger	Timeline
Age							
Below median	0.017	−0.018	−0.025+	−0.021	−0.021	−0.014	−0.031*
(n = 5,412)	(0.042)	(0.014)	(0.014)	(0.013)	(0.015)	(0.015)	(0.015)
Above Median	0.092+	−0.023	−0.014	−0.026+	0.019	0.024	0.009
(n = 5,271)	(0.047)	(0.014)	(0.014)	(0.014)	(0.014)	(0.015)	(0.014)
Difference	0.075	−0.005	0.011	−0.005	0.040*	0.038+	0.041*
	(0.063)	(0.020)	(0.020)	(0.019)	(0.020)	(0.021)	(0.021)
Gender							
Not male	0.032	−0.011	−0.016	−0.013	−0.013	0.004	−0.024
(n = 5,050)	(0.044)	(0.014)	(0.014)	(0.013)	(0.015)	(0.015)	(0.015)
Male	0.074+	−0.029*	−0.023	−0.034*	0.010	0.005	0.001
(n = 5,633)	(0.045)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
Difference	0.042	−0.018	−0.018	−0.018	0.023	0.023	0.023
	(0.063)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)
Supports governing party							
Not aligned	0.099**	−0.030*	−0.027*	−0.034**	−0.002	0.008	−0.010
(n = 7,570)	(0.038)	(0.012)	(0.012)	(0.011)	(0.012)	(0.012)	(0.012)
Aligned	−0.055	0.002	0.000	0.002	0.001	−0.003	−0.014
(n = 3,113)	(0.058)	(0.019)	(0.018)	(0.018)	(0.019)	(0.019)	(0.018)
Difference	−0.153*	0.032	0.032	0.032	0.002	0.002	0.002
	(0.069)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)
Digital literacy index							
Below median	0.058	−0.038**	−0.034*	−0.035*	−0.017	−0.007	−0.020
(n = 5,443)	(0.045)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
Above median	0.050	−0.003	−0.004	−0.013	0.016	0.017	−0.002
(n = 5,240)	(0.045)	(0.014)	(0.014)	(0.013)	(0.015)	(0.015)	(0.015)
Difference	−0.009	0.035+	0.030	0.022	0.033+	0.024	0.018
	(0.063)	(0.020)	(0.020)	(0.019)	(0.020)	(0.021)	(0.021)
Scientific knowledge index							
Below median	0.115**	−0.037**	−0.043**	−0.039**	−0.018	−0.004	−0.029*
(n = 5,677)	(0.043)	(0.014)	(0.014)	(0.013)	(0.014)	(0.014)	(0.014)
Above median	−0.015	−0.001	0.008	−0.007	0.018	0.014	0.009
(n = 5,006)	(0.046)	(0.015)	(0.014)	(0.014)	(0.015)	(0.015)	(0.015)
Difference	−0.130*	0.036+	0.051*	0.031	0.036+	0.018	0.038+
	(0.063)	(0.020)	(0.020)	(0.019)	(0.020)	(0.021)	(0.021)

Table S7. Heterogeneity in treatment effects under accuracy nudge by selected covariates. The sample is users in the evaluation stage, $n = 10{,}683$. Columns denote response measures, described in the note to Table S6 and in Subsection 3.1. Estimates are of treatment effects under the accuracy nudge, in contrast with the control condition. Estimates are produced from an augmented inverse probability weighted estimator, as described in Subsection 3.2, within specified subgroups. Under two-sided hypothesis tests: $^{+}\ p < 0.1$, $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$.

		False			True		
	Combined	Any sharing	Messenger	Timeline	Any sharing	Messenger	Timeline
Age							
Below median	−0.020	−0.021	−0.016	−0.012	−0.029+	−0.019	−0.040*
(n = 5,412)	(0.046)	(0.016)	(0.015)	(0.015)	(0.016)	(0.016)	(0.016)
Above Median	0.102+	−0.048**	−0.043**	−0.040**	−0.001	0.006	−0.003
(n = 5,271)	(0.053)	(0.016)	(0.016)	(0.015)	(0.016)	(0.016)	(0.016)
Difference	0.122+	−0.027	−0.028	−0.028	0.027	0.025	0.037
	(0.070)	(0.022)	(0.022)	(0.021)	(0.023)	(0.023)	(0.023)
Gender							
Not male	0.033	−0.031+	−0.027+	−0.025+	−0.035*	−0.018	−0.037*
(n = 5,050)	(0.049)	(0.016)	(0.016)	(0.015)	(0.017)	(0.017)	(0.017)
Male	0.047	−0.038*	−0.031*	−0.027+	0.002	0.003	−0.009
(n = 5,633)	(0.050)	(0.016)	(0.015)	(0.015)	(0.015)	(0.016)	(0.016)
Difference	0.015	−0.007	−0.007	−0.007	0.037	0.037	0.037
	(0.070)	(0.022)	(0.022)	(0.022)	(0.023)	(0.023)	(0.023)
Supports governing party							
Not aligned	0.087*	−0.041**	−0.031*	−0.040**	−0.012	−0.004	−0.021
(n = 7,570)	(0.042)	(0.013)	(0.013)	(0.012)	(0.014)	(0.014)	(0.014)
Aligned	−0.073	−0.019	−0.025	0.008	−0.024	−0.015	−0.026
(n = 3,113)	(0.065)	(0.020)	(0.020)	(0.020)	(0.020)	(0.021)	(0.021)
Difference	−0.159*	0.021	0.021	0.021	−0.012	−0.012	−0.012
	(0.077)	(0.024)	(0.024)	(0.024)	(0.025)	(0.025)	(0.025)
Digital literacy index							
Below median	0.042	−0.050**	−0.045**	−0.033*	−0.031*	−0.019	−0.033*
(n = 5,443)	(0.049)	(0.016)	(0.015)	(0.015)	(0.015)	(0.016)	(0.016)
Above median	0.039	−0.018	−0.013	−0.018	0.002	0.006	−0.011
(n = 5,240)	(0.050)	(0.016)	(0.015)	(0.015)	(0.017)	(0.017)	(0.017)
Difference	−0.003	0.031	0.032	0.015	0.033	0.025	0.021
	(0.070)	(0.022)	(0.022)	(0.021)	(0.023)	(0.023)	(0.023)
Scientific knowledge index							
Below median	0.054	−0.053***	−0.050***	−0.038*	−0.034*	−0.024	−0.037*
(n = 5,677)	(0.048)	(0.016)	(0.015)	(0.015)	(0.016)	(0.016)	(0.016)
Above median	0.025	−0.013	−0.005	−0.013	0.007	0.013	−0.006
(n = 5,006)	(0.051)	(0.016)	(0.016)	(0.015)	(0.016)	(0.016)	(0.016)
Difference	−0.029	0.040+	0.045*	0.024	0.041+	0.037	0.031
	(0.070)	(0.022)	(0.022)	(0.021)	(0.023)	(0.023)	(0.023)

Table S8. Heterogeneity in treatment effects under Facebook tips by selected covariates. The sample is users in the evaluation stage, $n = 10,683$. Columns denote response measures, described in the note to Table S6 and in Subsection 3.1. Estimates are of treatment effects under the Facebook tips, in contrast with the control condition. Estimates are produced from an augmented inverse probability weighted estimator, as described in Subsection 3.2, within specified subgroups. Under two-sided hypothesis tests: $^{+}\ p < 0.1$, $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$.