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

October 31, 2022

Author Contributions:

Competing Interest Statement:

Classification:

Keywords (3-5): Misinformation - covid - adaptive experiment

Abstract:

What online interventions are effective at reducing willingness to share health misinformation during an ongoing global pandemic? Using an adaptive experiment with Facebook users recruited on the platform in Kenya and Nigeria, we tested 40 combinations of interventions with the goal of improving sharing discernment– reducing intended sharing of false posts without adversely affecting true sharing– of COVID-19 posts. We estimate precise null effects of flagging misleading posts and including related articles, which are used by social media platforms. Instead, providing tips for spotting misinformation and nudging users to think about the accuracy of media content improves sharing discernment. Tips leads to effects equivalent to a nearly 8% reduction

add note with link to anonymized preanalysis plan

Paste the major and minor classification here. Dual classifications are permitted, but cannot be within the same major classification.

in intended sharing of false information, and nudging accuracy leads to a 4% reduction. We also find significant differences in response to these treatments across users with different characteristics, indicating these interventions affect outcomes through separate mechanisms. These low-cost scalable interventions may significantly improve the quality of information circulating online.

1 Main

Our contributions

1. Horse race among many interventions
2. Respondent works, not headline, and we learned two respondent that work especially well: FB vs accuracy. We can differentiate these to some extent.
3. Measuring sharing channel adds nuance (helps us better understand how treatments work, and for whom?)

LR: havent yet added this one in intro - will wait to see what we can say specifically

Alongside the outbreak of the novel coronavirus (SARS-CoV-2), much of the world's population also experienced an "infodemic"—the spread of misinformation related to the virus. Before effective vaccines were developed and widely available, people across the globe looked to alternative sources for 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). Understanding how to slow the spread of false "cures" may have life-saving consequences.

This paper focuses on these particularly dangerous pieces of COVID-19 misinformation – hoax "cures" – and tests numerous online interventions designed to curb sharing of these falsities on social media, while not adversely affecting the sharing of true information on COVID-19 prevention techniques. This study began in February 2021 before vaccines were widely available. Using targeted Facebook advertisements, we recruited a sample of social media users living in Kenya and Nigeria, two of the three largest Facebook markets in sub-Saharan Africa (World Population Review, 2022). Using a Facebook Messenger chatbot, we engaged participants in a survey experiment that recruited and kept these social media users on the platform where they would naturally engage with similar media posts. Participants answered survey questions and were randomized into different treatments delivered by the Messenger chatbot.

Our main outcome of interest is intended sharing discernment: whether respondents indicate wanting to share true but not false posts. We focus on sharing rather than belief, since exposure can further false narratives through resharing even if it doesn't affect an

individual's belief, and because exposure to COVID-19 misinformation can cause lower adoption of preventative behaviors (Bursztyn et al., 2022). In addition to measuring participant's intentions to share both true and false COVID-19 posts, we measured *how* they wanted to share each post – either publicly on their Timeline or privately on Messenger. We analyze/examine sharing intentions by post type/veracity as well as by sharing channel in order to understand users' sharing preferences at baseline and post treatment.

LR: maybe add another sentence here on specific response function in order to avoid going through setup in results section?

Towards the goal of understanding what works to improve sharing discernment we first ran a learning phase using a multi-factorial adaptive design to compare a set of 40 treatments against each other and control. We examine interventions delivered to both the individual user, such as tips for spotting fake news, training videos, and nudges; as well as treatments delivered on specific posts, such as flags or warning labels pinned on the article of interest. We differentiate between these types of interventions because they vary in their cost and scalability. This design allowed us to identify which respondent-level and headline-level treatments were most effective at reducing intentions to share false information without adversely affecting true sharing.[1] We differentiate between these types of interventions because they vary in their cost and scalability, specific post flags require resource-intensive fact checking sources to keep up to pace with the generation of new misinformation.

LR: I moved rest of this to discussion/-conclusion

Our adaptive algorithm optimized for treatments that improved a response function that weighed intention to share a false post twice as (negatively) as wanting to share a true post (positively).[2] The adaptive design sequentially assigned treatment probabilities to privilege assignment to the most effective interventions, and minimized assignment to ineffective or counter-productive interventions. [Still to add here: more specifics on response funct?, benefit to adaptivity...] Traditional randomized experiments are often limited in the number of interventions they can test due to power considerations, but our adaptive design allows us to sort through numerous interventions.

After identifying the two most successful headline-level and respondent-level interventions from the adaptive learning phase, we ran an evaluation phase to [identify the optimal policy/compare against each other and control]. With a new sample of Facebook users from Kenya and Nigeria in the evaluation phase, we tested factcheck labels and related articles headline-level treatments, and tips for spotting misinformation (Guess et al., 2020) and an accuracy nudge (Pennycook et al., 2021) for the respondent-level treatments. We find that the headline-level interventions do not perform better than control, and estimate precise null

I can add a lot of the design stuff in for the next edit, and then make consistent/remove duplication across here, results, and methods.

[1] Table S2 in the Supplementary Information (SI) describes all of the interventions we tested.

[2] In the next section we also show results of the interventions disaggregated by post type and sharing channel.

effects for these treatments. We do, however, see that tips and the accuracy nudge improve sharing discernment. [ADD specific findings re. heterogeneity / what we can say about distinguishing btw accuracy and tips...] By exploring heterogeneity in response to treatment we are able to say more specifically which types of users do best under which treatment/who should be targeted with each. The findings provide further evidence that these interventions are not perfect substitutes, perhaps because they operate through different mechanisms, and there are benefits to targeting users with different characteristics with tips over accuracy and vice versa.

[LR: will make last 4 paragraphs our contribution + lit review to serve that end]

LR: summarize limitations here and then put longer discussion in limitations subsection in discussion?

This study, like others focused on online misinformation, faces several limitations related to external validity. First, our goal is to identify interventions that are effective among the population of social media users in Kenya and Nigeria. We are limited, however, in our recruitment methods to engaging with those who clicked on our Facebook ads to participate in the study. Recruiting actual social media users on the platform is an improvement beyond 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 to the general population of Facebook users in these countries. Importantly, this study brings comparative data to this global question which is most often studied using samples recruited from Qualtrics, Lucid, and MTurk in North America.

Second, interacting with participants of a study and delivering interventions in the course of a survey experiment is an imperfect proxy for understanding how users would react to real interventions delivered on the platform. Though still artificial, our approach of delivering the survey and interventions through a Facebook Messenger chatbot provides greater realism than interventions delivered on other 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. The validity of our findings would be at risk if participants gleaned the intention of the study and adjusted their responses to match what they thought the researchers wanted to hear, rather than reflecting how they truly believed or wanted to behave. [TK (for now referenced on p.15): Later we presents tests that suggest our results are not wholly driven by experimenter demand bias.]

Finally, misinformation studies that focus on sharing behavior as the main outcome of interest are constrained by ethical considerations of not wanting to contribute to the ecosys-

tem of misinformation by allowing survey participants to *actually share* false posts. This study, like most others, instead use measures of sharing *intentions*. While scholars have found that intentions are correlated with online sharing behavior (Mosleh et al., 2020), measuring intentions rather than actual behavior remains a main limitation of scholarship in this area. In this study, we directly ask participants "Do you want to share this post on your timeline/on Messenger?," rather than phrasing it as a hypothetical question. We simultaneously told participants not to share the post now, but they would be able to do so at the end of the study. When we debriefed respondents at the end of the study, we told them which posts they saw were false and explained that was why they could not share those posts. We gave participants an opportunity to share the true posts they had said they wanted to share. A unique contribution of this study is that for each post (true and false), we asked participants if they wanted to share it on Timeline (public to their friends on Facebook) or on Messenger (a direct private message). We observe variation in sharing preferences by channel that suggest participants are discerning in their stated sharing intentions for true and false posts at baseline.

check this is correlated with true sharing intentions.

This paper investigates several important questions. Do social media users have different preferences for how they share true and false posts about COVID-19? Can we identify interventions that are effective at reducing the sharing of false information, without adversely affecting sharing of true information? We also explore whether there are benefits to policy targeting from two perspectives. First, we analyze which recipients should be targeted with interventions. Second, we observe whether particular interventions should be designated for certain types of users and not for others. This study is able to take a more comprehensive approach toward subgroup analysis by exploring who shares the most misinformation at baseline and who is most affected by treatment looking at covariates academic studies have found to be significant predictors, as well as characteristics platforms collect on users. With the goal of minimizing the spread of misinformation in the online information ecosystem, we quantify the best approach in this setting and offer lessons for other contexts.

A full page of limitations feels like a bit much in the first 4 substantive pages of the paper.

LR: will edit this once we nail down results section

2 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, respectively. We recruited social media users 18 years and over in these countries through targeted Facebook advertisements (see our advertisement in Figure S1 in the SI; for further details on targeted recruitment on Facebook, see Rosenzweig et al., 2020). 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 respondents: a "learning" stage, with 4,553 social media users, and an "evaluation" stage, with 10,681 social media users. In Supplementary Table S1 we report sample characteristics and comparison to 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. For each stimuli, users were separately asked whether they would like to share the stimuli through two channels: on Facebook Messenger and on their Facebook timeline.

Our pre-specified combined response measure is a weighted sum of times users said they would like to share true and misinformation stimuli over each channel. As our objective is to learn treatments that will decrease sharing of false information while not overly harming 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. More details on this measure are provided in Methods Section 4.1.

We also report results for both types of stimuli separately: we report the proportion of true and false stimuli respondents reported intending to share, across either Messenger or timeline, as well as disaggregated by sharing channel.

Learning and evaluation stages We designed the learning stage to compare a large range of treatment conditions, and learn which of them were most effective on our pre-specified combined response measure. We considered two classes of interventions: seven respondent-level interventions and four headline-level interventions. The respondent-level interventions included behavioral nudges and trainings targeted to the participants themselves: tips and trainings to spot false news (from Facebook, AfricaCheck, and a BBC video), an emotion suppression prompt, an accuracy nudge, and a pledge that participants took to keep their family and friends safe. The headline-level interventions were applied to the headlines or posts themselves: a flag for articles that had been fact checked by

third-party websites, links to further information, or accompanied by additional related articles or countering information from a validated source such as the WHO. Table S2 in the Supplementary Information (SI) describes all of the interventions we tested. We used a multi-factorial experimental design where each class of intervention was treated as a separate multi-level factor, with a baseline control condition.

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

This adaptive design allowed us to continue to learn which treatment was best, while reducing the probability that users were assigned to ineffective or harmful interventions. One methodological contribution of this study is to demonstrate the benefit of using adaptive experiments to learn more quickly and avoid sending poorly performing treatments - that may also backfire and increase sharing of misinformation - to more respondents, as conventional experiments with static treatment assignment probabilities do. 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 S3 in the SI illustrates the development of cumulative assignment under our assignment algorithm.

From the data in the learning stage, we selected two "best" treatments each from the respondent-level and headline-level classes. These were the two treatments associated with the highest estimated mean responses in each class separate from control, as measured on our combined response outcome. These treatments were the accuracy nudge and Facebook tips (respondent-level) and factcheck and related articles (headline-level); examples of each are presented in the Methods Section, Figure 4.

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 an optimal contextual policy, which assigned to each user the respondent-level treatment predicted to be best for them conditional on their covariate profile. We included only the two "best" respondent-level treatments in this optimal policy, as our results from the learning stage suggested that headline-level treatments at best improved only minimally over control (Figure S4 in the SI reports estimated response in the learning

stage). The following results are all based on data analyzed from the evaluation stage.[3]

Discernment under control Under the control condition, we see that users exhibit discernment in what types of stimuli they intend to share and through which channels. Overall, users report greater intentions to share true stimuli as compared to false stimuli on any channel (difference = 20.9 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 = 3.1 pp, SE = 0.7), but the reverse is true for false stimuli (difference = -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 among users. Understanding baseline heterogeneity in sharing behavior helps us to identify the greatest culprits of sharing misinformation. We focus on 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 pre-registered 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, older subjects, those aligned with the ruling party, participants with low digital literacy and those with low scientific knowledge have relatively lower valued outcomes on our combined response measure, indicating relatively higher false sharing intentions and/or lower true sharing intentions (see Table 1).

[3]To learn a contextual policy, we fit the augmented inverse probability weighted estimator described in Equation 2 to the learning data. From this model, we predicted responses for each covariate profile in the evaluation stage under counterfactual treatment conditions. The contextual policy assigned to each user the treatment associated with the highest predicted response, conditional on their covariates. In the main results here, we report results with respect to a policy learned on the false sharing outcome measure only, and include as possible treatments only the two "best" respondent-level treatments included in our evaluation.

Prior to collecting the evaluation data, we fit a contextual policy using the combined response measure, including a broader range of respondent-level treatments. However, as false sharing intentions are more responsive to treatment, the inclusion of true sharing intentions in the measure adds noise. We report outcomes for the policy learned on the combined response measure in Section S2.1 in the SI.

		False			True		
	Combined	Any sharing	Messenger	Timeline	Any sharing	Messenger	Timeline
Overall control mean							
	−0.391	0.442	0.395	0.369	0.651	0.561	0.593
	(0.027)	(0.009)	(0.008)	(0.008)	(0.008)	(0.009)	(0.009)
Age							
Below median	−0.484	0.457	0.413	0.374	0.632	0.548	0.562
(n = 5,412)	(0.035)	(0.012)	(0.012)	(0.011)	(0.012)	(0.013)	(0.012)
Above median	−0.295	0.425	0.380	0.361	0.670	0.575	0.624
(n = 5,271)	(0.040)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Difference	−0.189***	0.032+	0.033*	0.012	−0.038*	−0.028	−0.062***
	(0.054)	(0.017)	(0.017)	(0.016)	(0.017)	(0.018)	(0.017)
Gender							
Not male	−0.356	0.400	0.362	0.327	0.611	0.512	0.543
(n = 5,050)	(0.037)	(0.012)	(0.012)	(0.011)	(0.012)	(0.013)	(0.012)
Male	−0.422	0.478	0.427	0.404	0.687	0.605	0.637
(n = 5,633)	(0.038)	(0.012)	(0.012)	(0.012)	(0.011)	(0.012)	(0.012)
Difference	0.067	−0.077***	−0.077***	−0.077***	−0.076***	−0.076***	−0.076***
	(0.053)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)
Political allegiance							
Not aligned	−0.333	0.415	0.366	0.340	0.629	0.537	0.563
(n = 7,570)	(0.032)	(0.010)	(0.010)	(0.010)	(0.010)	(0.011)	(0.010)
Aligned	−0.532	0.504	0.469	0.434	0.704	0.621	0.664
(n = 3113)	(0.050)	(0.016)	(0.016)	(0.015)	(0.015)	(0.016)	(0.016)
Difference	0.199***	−0.089***	−0.089***	−0.089***	−0.075***	−0.075***	−0.075***
	(0.059)	(0.019)	(0.019)	(0.019)	(0.018)	(0.018)	(0.018)
Digital literacy index							
Below median	−0.536	0.495	0.448	0.421	0.674	0.587	0.620
(n = 5,443)	(0.038)	(0.012)	(0.012)	(0.012)	(0.011)	(0.012)	(0.012)
Above median	−0.240	0.385	0.343	0.313	0.627	0.535	0.564
(n = 5,240)	(0.037)	(0.012)	(0.012)	(0.011)	(0.012)	(0.013)	(0.013)
Difference	−0.296***	0.110***	0.106***	0.108***	0.048**	0.052**	0.056**
	(0.053)	(0.017)	(0.017)	(0.016)	(0.017)	(0.018)	(0.017)
Scientific knowledge index							
Below median	−0.445	0.458	0.414	0.385	0.658	0.563	0.597
(n = 5,677)	(0.036)	(0.012)	(0.012)	(0.011)	(0.012)	(0.012)	(0.012)
Above median	−0.330	0.422	0.377	0.347	0.643	0.559	0.587
(n = 5,006)	(0.040)	(0.013)	(0.012)	(0.012)	(0.012)	(0.013)	(0.013)
Difference	−0.115*	0.036*	0.036*	0.038*	0.015	0.004	0.010
	(0.054)	(0.017)	(0.017)	(0.016)	(0.017)	(0.018)	(0.017)

Table 1. Heterogeneity in response under the control condition by selected covariates. Estimates are of mean response under the control condition, and are produced from an augmented inverse probability weighted estimator, as described in Section 4.1, within specified subgroups. For contrasts only, under two-sided hypothesis tests: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

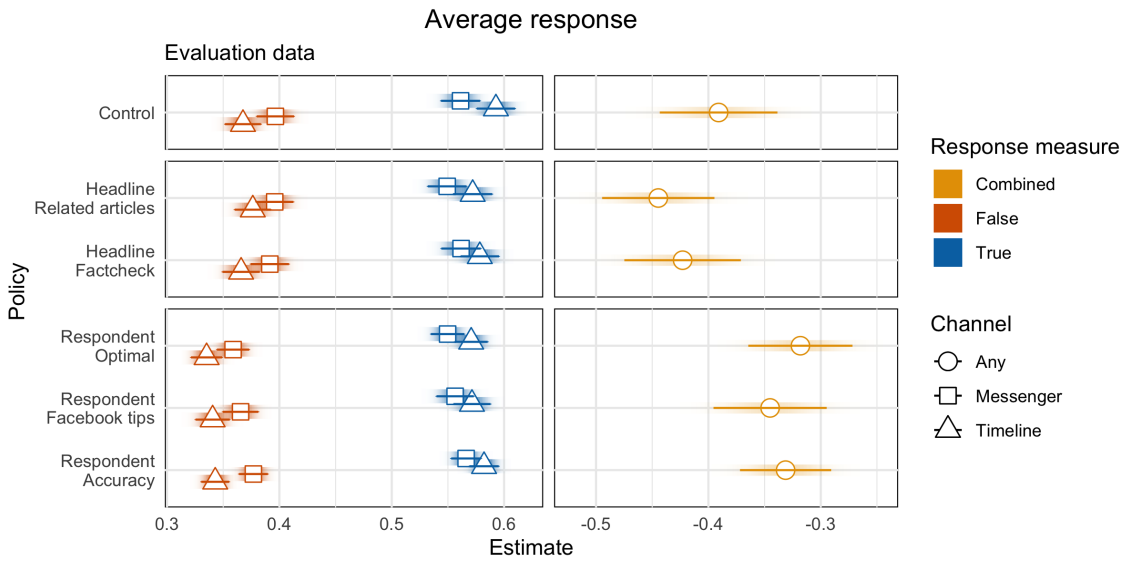

Figure 1. Response estimates. Response measures are average intention to share true and false stimuli over either channel, and a combined response measure, reported in Section 4.1. Estimates are produced from an augmented inverse probability weighted estimator, as described in Section 4.1.

Main treatment effects Table 2 shows tests of our pre-specified comparisons of each evaluated treatment condition against the control. We report results for the evaluation phase under the pre-registered combined response function, as well as disaggregated by any intention to share false and true stimuli by either channel, and false and true sharing intentions by each channel. 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, and positive or neutral treatment effects on true sharing; we pre-registered one-sided hypothesis tests with respect to these treatment effects.

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 factcheck treatment is associated with a decrease of 0.5 pp (SE = 1.2) in false sharing intentions as compared to control; the effect would need to be nearly four times as large with the same

degree of uncertainty for the confidence interval to exclude zero.

The respondent-level treatments, however, are effective. The Facebook tips and accuracy nudge treatments increase the combined response measure by 4.6 pp (SE = 3.5) and 6.0 pp (SE = 3.2) relative to control, respectively. These effects are driven by decreases in false sharing of 3.3 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	−0.032	−0.005	−0.004	−0.005	−0.003	−0.001	−0.012
	(0.036)	(0.012)	(0.011)	(0.011)	(0.011)	(0.012)	(0.012)
Related articles	−0.054	0.008	0.002	0.009	−0.019	−0.012	−0.021
	(0.035)	(0.012)	(0.011)	(0.011)	(0.011)	(0.012)	(0.012)
Respondent treatment effects							
Accuracy	0.060*	−0.020*	−0.018*	−0.026**	−0.001	0.005	−0.011
	(0.032)	(0.010)	(0.010)	(0.009)	(0.010)	(0.010)	(0.010)
Facebook tips	0.046+	−0.033**	−0.029**	−0.030**	−0.016	−0.005	−0.022
	(0.035)	(0.011)	(0.011)	(0.010)	(0.011)	(0.012)	(0.011)
Optimal	0.073*	−0.039***	−0.038***	−0.032***	−0.019	−0.012	−0.022
	(0.034)	(0.011)	(0.010)	(0.010)	(0.011)	(0.011)	(0.011)
Control mean	−0.391	0.442	0.395	0.369	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 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 Section 4.1. $n = 10,681$. For contrasts only, under one-sided hypothesis tests, as pre-specified in pre-registration: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Heterogeneity in best policy While both the Facebook tips and accuracy nudge are effective, we also observe differences in how users respond to these two treatments. Figure 2 shows differences in average response under the accuracy nudge as compared to Facebook tips, if we were to assign the accuracy nudge according to a prioritization rule instead of at random, following the approach presented in Yadlowsky et al. (2021). Here the prioritization

rule is assigned by fitting a causal forest on the learning data, predicting response under the model on the evaluation data, and ordering based on predicted differences. We fit a separate model on the evaluation data to estimate the efficacy of the evaluation rule. As false sharing intentions are more responsive to treatment, we focus here on a rule based on propensity to share false information. We can see, for example, that if we were limited to assigning the accuracy nudge to only 40 percent of the population, false sharing intentions would be 4.4 pp lower (SE = 1.5) 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 2, is -3.8 pp (SE = 1.3), using the targeting operator characteristic curve.

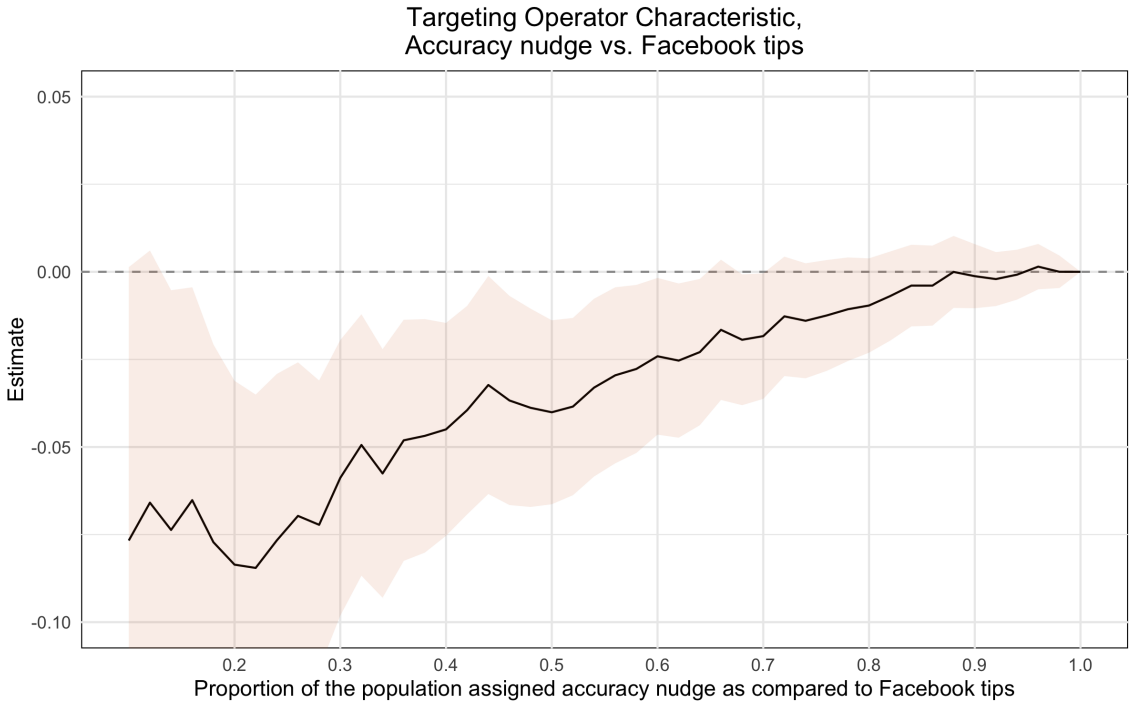

Figure 2. Targeting operator characteristic curve, comparing the accuracy nudge and Facebook tips. The outcome measure is the difference in proportion of false stimuli participants reported wanting to share, either as a Facebook post or privately in Facebook Messenger, between the accuracy nudge and 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.

To evaluate the overall effect of targeting, we consider the causal forest model learned for

prioritization rule above as a contextual policy: if the predicted difference between the accuracy nudge and the Facebook tips treatment is negative (indicating that the accuracy nudge is more effective at decreasing false sharing), our policy assigns the accuracy nudge; otherwise the Facebook tips treatment is assigned. Again, as the policy is learned on the learning data and evaluated on the evaluation data, we avoid concerns with over-fitting.

When the model is applied to the evaluation data, our optimal contextual policy assigns 43.5% of participants to Facebook tips, which is the best uniform policy for decreasing sharing of false stimuli. Optimally assigning these treatments, we achieve a treatment effect of -3.9 percentage points (SE = 1.1) in decreasing false sharing intentions. We saw in Table 2 that we achieve larger magnitude treatment effects in decreasing false sharing intentions through our contextual policy as compared to either the accuracy nudge or the Facebook tips treatments assigned uniformly (differences of -2.0 pp, SE = 0.8 and -0.8 pp, SE = 0.8, one-sided p-values of 0.009 and 0.150, respectively).

In Table 3, we see that our contextual policy learned on the learning data is appropriately assigning participants to the respective respondent-level conditions: participants for whom assignment under the learned optimal policy is Facebook tips on average intend to share false information at lower rates under the Facebook tips treatment as compared to the accuracy nudge (difference of 3.8, SE = 1.4); the reverse is true directionally for the participants assigned the accuracy nudge under the learned optimal policy (difference of -1.6, SE = 1.5).

		False			True		
	Combined	Any sharing	Messenger	Timeline	Any sharing	Messenger	Timeline
Optimal assignment == Accuracy nudge (n = 4,967)							
Accuracy	−0.162	0.396	0.349	0.341	0.681	0.588	0.637
	(0.031)	(0.010)	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)
Facebook tips	−0.220	0.412	0.363	0.352	0.687	0.602	0.639
	(0.038)	(0.012)	(0.012)	(0.011)	(0.013)	(0.013)	(0.013)
Difference	0.058	−0.016	−0.014	−0.012	−0.006	−0.014	−0.001
	(0.050)	(0.015)	(0.015)	(0.015)	(0.016)	(0.016)	(0.016)
Optimal assignment == Facebook tips (n = 5,716)							
Accuracy	−0.478	0.445	0.401	0.345	0.622	0.547	0.534
	(0.027)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Facebook tips	−0.453	0.407	0.367	0.330	0.589	0.516	0.513
	(0.035)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Difference	−0.025	0.038**	0.034*	0.014	0.033*	0.031*	0.021
	(0.044)	(0.014)	(0.014)	(0.014)	(0.015)	(0.015)	(0.015)
Grand difference							
	0.083	0.054*	0.048*	0.026	0.039+	0.045*	0.023
	(0.066)	(0.021)	(0.021)	(0.020)	(0.022)	(0.022)	(0.021)

Table 3. Response under counterfactual uniform respondent treatment conditions, by contextual policy assignment. Estimates are of mean response under the two respondent-level treatments. Estimates are produced from an augmented inverse probability weighted estimator, as described in Section 4.1, 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 an accuracy nudge both operate along the mechanism of increasing attention to accuracy, as suggested for the accuracy nudge by Pennycook et al. (2019), or rather whether the 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 3.8 pp, SE = 1.4) suggests, however, that there are differences in how users respond to these treatments, and different types of people respond differently to each treatnent.

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 46.5% of participants assigned to the accuracy nudge are, on average, more digitally literate, more likely to have more scientific knowledge, more likely to be male, older, and directionally more likely to be aligned with the governing political party than those assigned to the

Facebook tips treatment (see Figure 3).

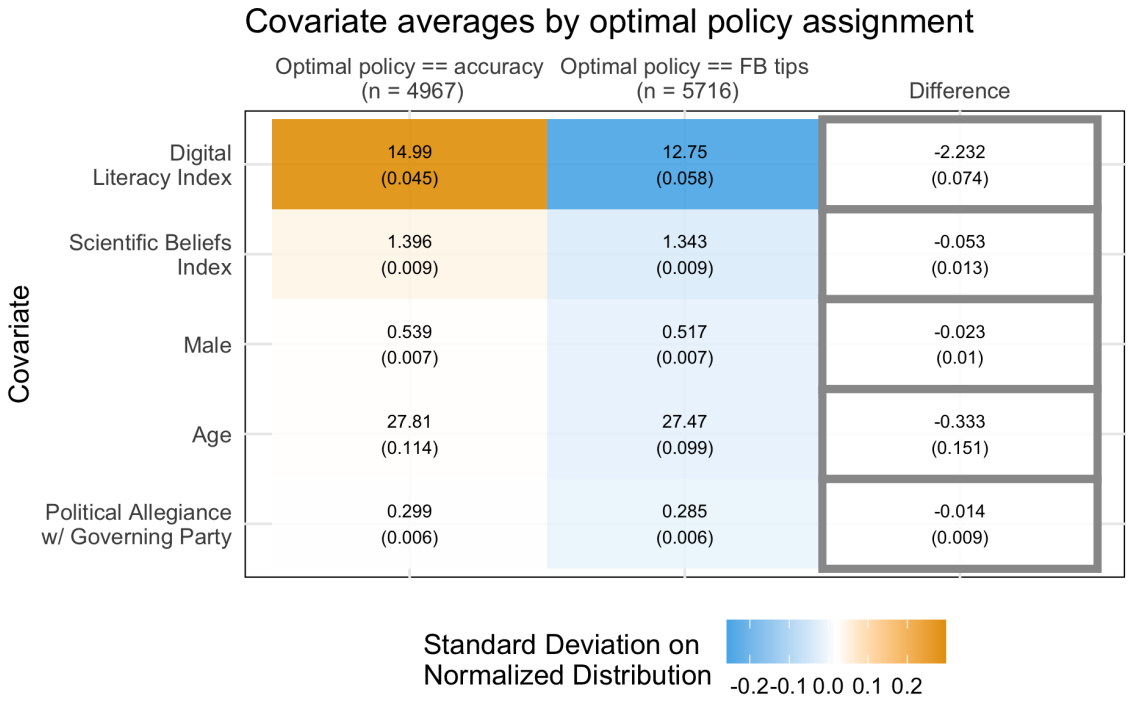

Figure 3. Selected covariate means between participants assigned to the accuracy nudge as compared to participants assigned to Facebook tips under the contextual policy. Covariates are ordered by size of standardized deviation between the two groups.

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 = 0.9), but it also directionally reduces true sharing intentions (difference of -1.5 pp, SE = 1.0); this 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.

The Facebook tips treatment and the accuracy nudge are both effective at moving false sharing intentions on the timeline (-3.0 pp, SE = 1.0; -2.6 pp, SE = 0.9 respectively) and on Messenger (-2.9 pp, SE = 1.1; -1.8 pp, SE = 1.0 respectively) relative to control. The

Facebook tips treatment is directionally relatively more effective at also reducing false sharing intentions on Messenger (difference of -0.9, 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 that respondents share false stimuli at all is that they are not able to discern between true and false stimuli. As noted, we do see evidence of discernment under control, where respondents 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 treatment, 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 respondents 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 should inform relative effects on the channel by which respondents share false 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., 2022b), but they may still believe that the posts may be of interest or value to share with individual contacts. If increased attention to the 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.

Heterogeneous response and treatment effects We find 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. 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 (-4.3, SE = 1.4), men (-3.3, SE = 1.3), and participants with low scientific knowledge (-4.5, SE =1.3). (See Table 4.) 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.

This 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 1.0 pp, SE = 1.3 for digital literacy; 1.4 pp, SE = 1.3 for scientific knowledge) (see Tables S6 and S5 in the SI). One possible explanation is that lower digital literacy and lower scientific knowledge users may be less able to differentiate false from true stimuli, and so the larger effects on these groups would be consistent with these users learning how to better evaluate stimuli under the Facebook tips treatment. In other words, the first order concern for these users is to equip them with the techniques and skills to identify misinformation, whereas for others with higher levels of digital skills and scientific knowledge interventions reminding users to focus on accuracy are more effective.

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 1.1 pp, SE = 1.2 for digital literacy; 0.9 pp, SE = 1.1 for scientific knowledge), whereas for the accuracy nudge, effects on timeline as compared to messenger sharing are comparable for these groups (difference of -0.3 pp, SE = 1.1 for digital literacy; 0.1 pp, SE = 1.1 for scientific knowledge). The overall larger effects on private Messenger sharing for Facebook tips as compared to the accuracy nudge is 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	0.002	−0.019	−0.021	−0.017	−0.026+	−0.020	−0.034*
(n = 5,412)	(0.039)	(0.013)	(0.013)	(0.012)	(0.014)	(0.014)	(0.014)
Above median	0.104*	−0.032*	−0.030*	−0.035**	0.009	0.020	0.003
(n = 5,271)	(0.045)	(0.014)	(0.013)	(0.013)	(0.013)	(0.014)	(0.014)
Difference	−0.102+	0.013	0.009	0.019	−0.035+	−0.040*	−0.037+
	(0.060)	(0.019)	(0.018)	(0.018)	(0.019)	(0.020)	(0.019)
Gender							
Not male	0.038	−0.020	−0.024+	−0.020	−0.025+	−0.003	−0.030*
(n = 5,050)	(0.042)	(0.014)	(0.013)	(0.012)	(0.014)	(0.014)	(0.014)
Male	0.066	−0.030*	−0.026*	−0.031*	0.006	0.003	−0.004
(n = 5,633)	(0.042)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
Difference	−0.029	0.011	0.003	0.011	−0.031	−0.007	−0.026
	(0.059)	(0.019)	(0.018)	(0.018)	(0.019)	(0.020)	(0.019)
Political allegiance							
Not aligned	0.095**	−0.034**	−0.030**	−0.039***	−0.008	0.003	−0.014
(n = 7,570)	(0.035)	(0.011)	(0.011)	(0.011)	(0.011)	(0.012)	(0.012)
Aligned	−0.049	−0.005	−0.013	0.005	−0.011	−0.008	−0.022
(n = 3,113)	(0.055)	(0.018)	(0.017)	(0.017)	(0.017)	(0.018)	(0.017)
Difference	0.144*	−0.029	−0.017	−0.044*	0.004	0.011	0.008
	(0.065)	(0.021)	(0.020)	(0.020)	(0.021)	(0.021)	(0.021)
Digital literacy index							
Below median	0.056	−0.041**	−0.039**	−0.035**	−0.025+	−0.010	−0.026*
(n = 5,443)	(0.042)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
Above median	0.049	−0.009	−0.011	−0.016	0.008	0.011	−0.005
(n = 5,240)	(0.042)	(0.014)	(0.013)	(0.012)	(0.014)	(0.014)	(0.014)
Difference	0.007	−0.033+	−0.028	−0.019	−0.033+	−0.021	−0.021
	(0.060)	(0.019)	(0.018)	(0.018)	(0.019)	(0.020)	(0.019)
Scientific knowledge index							
Below median	0.094*	−0.043**	−0.047***	−0.042***	−0.027*	−0.012	−0.033*
(n = 5,677)	(0.040)	(0.013)	(0.013)	(0.012)	(0.013)	(0.014)	(0.013)
Above median	0.006	−0.006	−0.001	−0.008	0.012	0.014	0.004
(n = 5,006)	(0.044)	(0.014)	(0.013)	(0.013)	(0.014)	(0.014)	(0.014)
Difference	0.087	−0.037+	−0.045*	−0.034+	−0.039*	−0.026	−0.037+
	(0.060)	(0.019)	(0.018)	(0.018)	(0.019)	(0.020)	(0.019)

Table 4. Heterogeneity in treatment effects under averaged respondent-level treatments by selected covariates. 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 Section 4.1, within specified subgroups. Under two-sided hypothesis tests: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

3 Discussion

We find both the accuracy nudge and Facebook tips are effective at curbing intentions to share misinformation. These treatments relate to two theories about why people are susceptible to misinformation.[4] The first from cognitive science suggests that people consume social media content quickly, react intuitively and do not stop to think about whether something is true or false. This reasoning suggests that people need to be reminded or "nudged" to consider the accuracy of posts, otherwise it may not be something they consider before sharing a post (Pennycook and Rand, 2021; Pennycook et al., 2021). A second rationale suggests that people simply do not know how to identify misinformation or lack the information to be able to do so (Ecker et al., 2022). This *information deficit model* prescribes providing training or tips to equip individuals to be able to spot misinformation in their news feed. Importantly, these are not mutually exclusive theories and people may suffer from both challenges—but whether one intervention is more successful, on average, than the other is important to understand as well as for which types of people one prescription may be better than the other.

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. Pennycook et al. (2021) provide evidence that increased attention to the accuracy of articles is the mechanism driving the efficacy of the accuracy nudge.

We also evaluated two headline-level treatments. We tested a factcheck intervention that has been used by several platforms and adds "disputed" flags to false posts. We also tested the related articles intervention that Facebook has used in the past—providing links to related articles under misleading or false posts (Ghosh, 2017).

Warning labels on posts have been found to be effective at helping users identify misinformation (Clayton et al., 2020) and reduce individual's willingness to share fake-news headlines (Mena, 2020) in the context of political information in respondent samples from the Global North. For COVID-19 information, Kreps and Kriner (2020) find the

[4]There are of course additional theories to explain why people are susceptible to or share misinformation—for instance for identity-based or ideological reasons (Nyhan and Reifler, 2010), or because something is funny or interesting if true (Altay et al., 2022a).

effectiveness of these tags to depend on context, specifically that they worked for only one out of three false COVID headlines tested. Recent evidence suggests fact checks can improve discernment (belief) among samples from Africa, Latin America, and the UK (Porter and Wood, 2021). Brashier et al. (2021) also find that the timing of fact checking matters—specifically that debunking misinformation after the headline is shown was more successful than contemporaneous tags, but do not test how these flags and warnings affect sharing behavior.

Interventions that flag specific posts rely on resource intensive fact-checking sources to verify the veracity of individual posts before applying such labels. Most posts are only flagged after they have circulated online and become popular, having the opportunity to inflict damage before factchecks and labels can be applied. A strategy of fact-checking and adding labels requires keeping up with new misinformation being generated. General interventions delivered to users while they are on a particular platform, on the other hand, are less resource intensive, can be applied at any time, and are much more easily delivered en masse.

cut? or move to discussion? or make 1 sentence in prior par?

This study, like others of its kind, has limitations. Importantly, the data come from a survey experiment, rather than a field experiment. However, the design of the study offers enhanced realism for participants beyond the standard approach of recruitment of samples through survey firms, and implementation of surveys on web browser-based platforms. Instead, we recruit social media users on the Facebook platform itself, where they would normally come into contact with online (mis)information about COVID-19. We keep participants on the platform interacting with a Messenger bot, which may feel somewhat more naturalistic than answering survey questions using another software, but is still an experimenter-controlled environment. This control facilitated more straightforward measurement, and also reduces ethical concerns about the possibility of the experiment facilitating the spread of COVID misinformation during a global pandemic.

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 many diverse contexts. This study provides evidence that such interventions are more effective than many others often tested by academics 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. Policymakers and platforms may also consider targeting interventions to those prone to share misinformation and not waste resources or risk a worse user experience by *not* directing such interventions to groups for whom they are ineffective. Overall, these policies delivered to participants are

whether we include here/intro depends on how much of contribution we see this as
- I see it mostly as a nice point to make: we were able to evaluate on same sample/-time/measures a ton of treatment across these 2 types
- we see respondent-level more effective –¿ might be worth investing more in finding successes of these types (bc cheaper + easier to scale). so I

much more cost effective than headline-specific interventions that require time and effort from human or AI fact checkers.

Add more about what we can say re. differences in FB tips and accuracy

Experimenter demand effects To attempt to reduce 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 the survey objectives, and treatment effects to be largest for this group. Instead, we see the reverse is true. The variation in treatment effects by channel 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.

4 Methods

4.1 Data and recruitment

Our sample is recruited from Facebook users in Kenya and Nigeria. Kenya and Nigeria represent two of Facebook's top three largest user bashes in sub-Saharan Africa (ITNews, 2016), with a combined user base of 30-35 million users ages 18 years and older.[5] We used targeted Facebook advertisements to improve balance on age and gender. Users who clicked on our ads offering airtime for taking a survey (see Figure S1) then started a conversation with our page's Messenger chatbot. Participants who completed the survey received compensation in the form of mobile phone airtime (equivalent to about $0.50) sent to their phone.

Outcome measures Each participant saw four post-treatment stimuli, two true and two false in a random order. For each stimuli, we asked respondents 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

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

AFP, Poynter, and AfricaCheck websites lists of misinformation that had appeared online and which was fact-checked in Kenya and Nigeria since the start of the pandemic.

Our outcomes include the combined response measure We control for strata of pre-test response (Davidian et al., 2005) and use an index of repeated measures (Broockman et al., 2017) to improve the efficiency of effect estimation.

Treatments We considered two types of treatments: respondent-level interventions that included behavioral nudges and trainings targeted to the participants themselves; and headline-level interventions that were applied to stimuli. In the evaluation stage, we tested the two of each type of intervention against control, along with a contextual policy composed of the two respondent-level treatments.

The selected treatments were the accuracy nudge and Facebook tips (respondent-level) and factcheck 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 treatment is presented in the SI Section S1.3. The factcheck treatment included a warning label on false stimuli, modeled on one used by Facebook for its third-party factchecking 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 4.

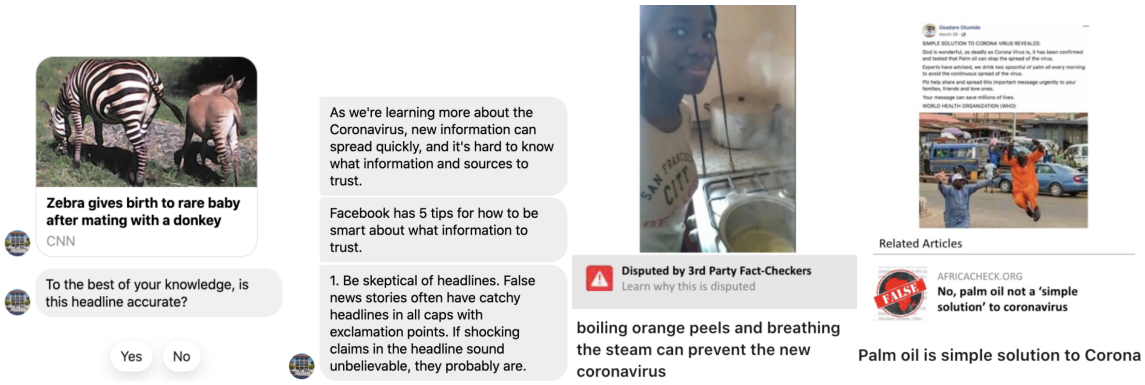

Figure 4. Respondent- and headline-level treatments tested in the evaluation phase.

Empirical strategy For both the learning and the evaluation stages of our study, we conduct estimation accounting both for unequal treatment assignment probabilities, and adjustment 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). To account for non-normality of the estimator on adaptively collected data, we use adaptive weights, described in Zhan et al. (2021a).

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 $e_i(w) := \Pr[W_i = w | X_i = x]$. We estimate the conditional means model using a random forest as implemented by the `grf` page in R statistical software (Tibshirani et al., 2020).

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 S3.

5 Acknowledgements

We received advertising credits for this study from Facebook Health. For exceptional research assistance, we thank James (Zelin) Li, Ricardo Ruiz, and Undral Byambadalai. 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 Becker Friedman Institute.

References

Altay, S., de Araujo, E., and Mercier, H. (2022a). "if this account is true, it is most enormously wonderful": Interestingness-if-true and the sharing of true and false news. *Digital Journalism*, 10(3):373–394.

Altay, S., Hacquin, A.-S., and Mercier, H. (2022b). 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.

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.

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.

Brennen, J. S., Simon, F. M., Howard, P. N., and Nielsen, R. K. (2020). Types, sources, and claims of covid-19 misinformation. *Reuters Institute*.

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.

Bursztyn, L., Rao, A., Roth, C., and Yanagizawa-Drott, D. (2022). Opinions as facts. Technical report, ECONtribute Discussion Paper.

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

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.

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.

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.

Kreps, S. E. and Kriner, D. (2020). Medical misinformation in the covid-19 pandemic. *Available at SSRN 3624510*.

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

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

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.

Murphy, J., Vallières, F., Bentall, R. P., Shevlin, M., McBride, O., Hartman, T. K., McKay, R., Bennett, K., Mason, L., Gibson-Miller, J., et al. (2021). Psychological characteristics associated with covid-19 vaccine hesitancy and resistance in ireland and the united kingdom. *Nature communications*, 12(1):1–15.

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. (2021). The psychology of fake news. *Trends in cognitive sciences*, 25(5):388–402.

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*.

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

S1.1 Recruitment

Placeholder

Table S1. Demographics

Add in table about sample representativeness, compare to Afrobarometer

Figure S1. Advertising image used for recruitment.

S1.2 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

S1.3 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. 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 to be effective at improving discernment of online political information. Emotions 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 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]citepcosta2018walking. 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 S2. Full list of treatments run during the learning phase.

S1.3.1 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.

S1.4 Covariates

In all analyses, we include the pre-test 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) $\times$ (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 S3. Covariates and response options

6

S1.5 Response measurement

We are primarily interested in decreasing sharing of harmful false information about COVID-19 cures and treatments, however, we simultaneously wish to limit any negative impacts 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?

[6]Trust in science has been found to be a particularly strong predictor of belief in COVID misinformation/conspiracy theories (Murphy et al., 2021).

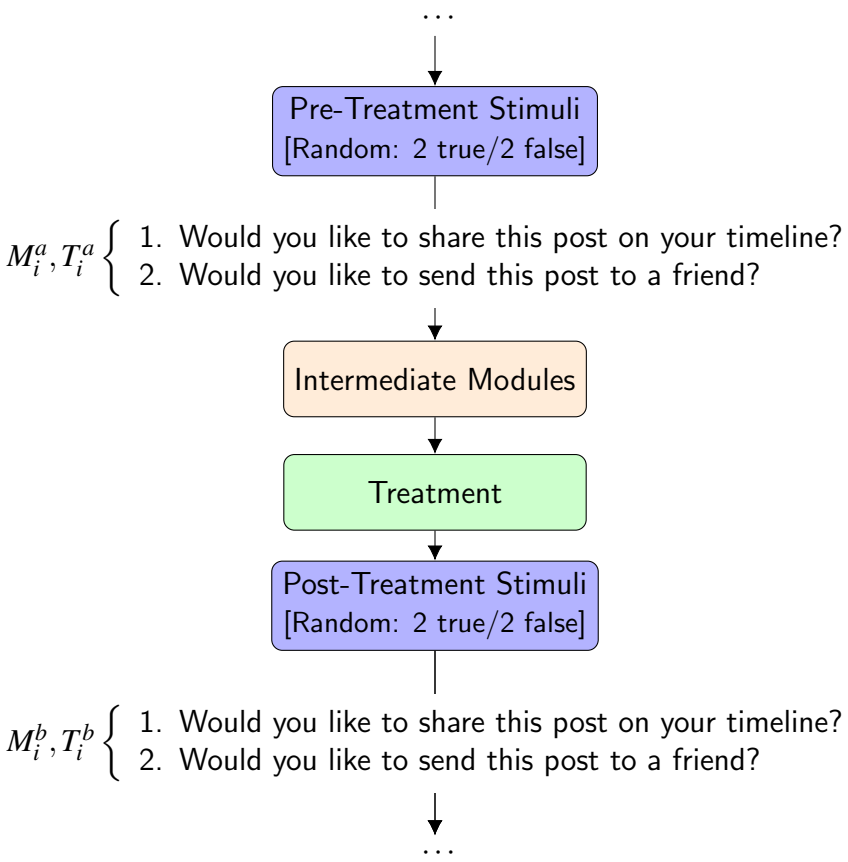

Figure S2. Survey flow.

By using a pre-test / post-test 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 (see the Supporting Information for the set of posts we used). 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 are then 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 pre-treatment. 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 response 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 pre-test responses to the *misinformation* stimuli and let T_i^a be the sum of respondent i's pre-test responses to the *true* informational stimuli. M_i^b and T_i^b are the respective post-treatment responses. Then $M_i^a, T_i^a, M_i^b, T_i^b \in \{0,1,2,3,4\}$.

We control for strata of pre-test responses in our analyses. We formalize our response function in terms of post-test measures:

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

This response function is the metric that we optimize for in our adaptive algorithm. Table S4 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 S4. Combined response measure.

S2 Additional results

S2.1 Learning stage

Revise wording for this section based on Susan's feedback

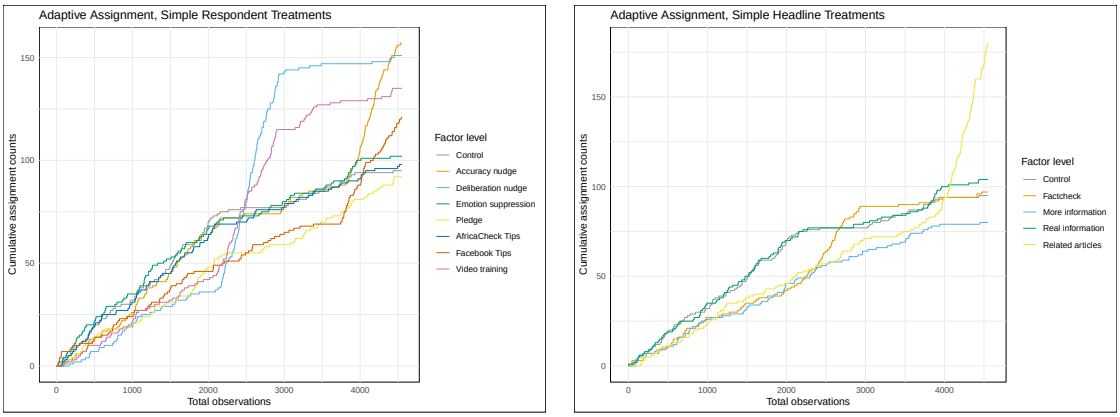

Figure S3. Cumulative treatment assignment during the learning phase for headline (left panel) and respondent (right panel) interventions. While the full design allows for all factor combinations, these plots illustrate assignment using the "pure" version of each factor, i.e., when the other factor is at the baseline control condition.

show something about the bayesian posteriors too?

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.

[TK]

Figure S4. Learning stage estimates grid

Add this grid

S2.2 Evaluation stage

		False			True		
	Combined	Any sharing	Messenger	Timeline	Any sharing	Messenger	Timeline
Age							
Below median	0.022	−0.018	−0.024+	−0.021	−0.022	−0.018	−0.028*
(n = 5,412)	(0.042)	(0.014)	(0.014)	(0.013)	(0.015)	(0.015)	(0.014)
Above median	0.099*	−0.020	−0.015	−0.028*	0.020	0.028+	0.008
(n = 5,271)	(0.047)	(0.014)	(0.014)	(0.014)	(0.014)	(0.015)	(0.014)
Difference	0.077	−0.001	0.008	−0.007	0.041*	0.046*	0.036+
	(0.064)	(0.020)	(0.020)	(0.019)	(0.020)	(0.021)	(0.020)
Gender							
Not male	0.037	−0.010	−0.017	−0.014	−0.014	0.007	−0.023
(n = 5,050)	(0.045)	(0.015)	(0.014)	(0.013)	(0.015)	(0.015)	(0.015)
Male	0.079+	−0.026+	−0.022	−0.035*	0.010	0.003	0.000
(n = 5,633)	(0.045)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
Difference	0.042	−0.016	−0.016	−0.016	0.024	0.024	0.024
	(0.063)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)
Political allegiance							
Not aligned	0.102**	−0.029*	−0.028*	−0.035**	−0.002	0.008	−0.008
(n = 7,570)	(0.038)	(0.012)	(0.012)	(0.011)	(0.012)	(0.012)	(0.012)
Aligned	−0.043	0.006	0.001	0.001	0.001	−0.003	−0.017
(n = 3,113)	(0.058)	(0.019)	(0.018)	(0.018)	(0.019)	(0.019)	(0.018)
Difference	−0.145*	0.035	0.035	0.035	0.004	0.004	0.004
	(0.070)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)
Digital literacy index							
Below median	0.065	−0.036*	−0.033*	−0.036**	−0.017	−0.006	−0.020
(n = 5,443)	(0.045)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
Above median	0.054	−0.001	−0.006	−0.013	0.016	0.016	−0.001
(n = 5,240)	(0.045)	(0.014)	(0.014)	(0.013)	(0.015)	(0.015)	(0.015)
Difference	−0.010	0.036+	0.027	0.023	0.033	0.022	0.019
	(0.064)	(0.020)	(0.020)	(0.019)	(0.020)	(0.021)	(0.020)
Scientific knowledge index							
Below median	0.126**	−0.036*	−0.043**	−0.042**	−0.018	−0.004	−0.030*
(n = 5,677)	(0.043)	(0.014)	(0.014)	(0.013)	(0.014)	(0.014)	(0.014)
Above median	−0.015	0.000	0.007	−0.005	0.018	0.015	0.012
(n = 5,006)	(0.047)	(0.015)	(0.014)	(0.014)	(0.015)	(0.015)	(0.015)
Difference	−0.141*	0.036+	0.050*	0.037+	0.037+	0.018	0.042*
	(0.064)	(0.020)	(0.020)	(0.019)	(0.020)	(0.021)	(0.020)

Table S5. Heterogeneity in treatment effects under accuracy nudge by selected covariates. 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 Section 4.1, 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	−0.019	−0.018	−0.012	−0.031+	−0.022	−0.040*
(n = 5,412)	(0.046)	(0.016)	(0.015)	(0.014)	(0.016)	(0.016)	(0.016)
Above median	0.110*	−0.044**	−0.045**	−0.042**	−0.002	0.013	−0.003
(n = 5,271)	(0.053)	(0.016)	(0.016)	(0.015)	(0.016)	(0.016)	(0.016)
Difference	0.127+	−0.025	−0.027	−0.030	0.029	0.034	0.037
	(0.070)	(0.022)	(0.022)	(0.021)	(0.023)	(0.023)	(0.023)
Gender							
Not male	0.038	−0.029+	−0.030+	−0.027+	−0.036*	−0.014	−0.037*
(n = 5,050)	(0.050)	(0.016)	(0.016)	(0.015)	(0.017)	(0.017)	(0.017)
Male	0.053	−0.034*	−0.031*	−0.027+	0.001	0.003	−0.007
(n = 5,633)	(0.050)	(0.015)	(0.015)	(0.015)	(0.015)	(0.016)	(0.016)
Difference	0.016	−0.006	−0.006	−0.006	0.037	0.037	0.037
	(0.070)	(0.022)	(0.022)	(0.022)	(0.023)	(0.023)	(0.023)
Political allegiance							
Not aligned	0.087*	−0.039**	−0.032*	−0.042***	−0.013	−0.001	−0.019
(n = 7,570)	(0.042)	(0.013)	(0.013)	(0.012)	(0.014)	(0.014)	(0.014)
Aligned	−0.055	−0.015	−0.027	0.010	−0.024	−0.013	−0.027
(n = 3,113)	(0.065)	(0.021)	(0.020)	(0.020)	(0.020)	(0.021)	(0.021)
Difference	−0.143+	0.024	0.024	0.024	−0.011	−0.011	−0.011
	(0.078)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)
Digital literacy index							
Below median	0.048	−0.046**	−0.046**	−0.034*	−0.032*	−0.015	−0.033*
(n = 5,443)	(0.050)	(0.016)	(0.015)	(0.015)	(0.015)	(0.016)	(0.015)
Above median	0.043	−0.017	−0.016	−0.019	0.000	0.006	−0.009
(n = 5,240)	(0.050)	(0.016)	(0.016)	(0.014)	(0.017)	(0.017)	(0.017)
Difference	−0.005	0.030	0.030	0.015	0.032	0.020	0.024
	(0.070)	(0.022)	(0.022)	(0.021)	(0.023)	(0.023)	(0.023)
Scientific knowledge index							
Below median	0.061	−0.050**	−0.050**	−0.041**	−0.036*	−0.021	−0.036*
(n = 5,677)	(0.048)	(0.016)	(0.015)	(0.015)	(0.016)	(0.016)	(0.016)
Above median	0.028	−0.011	−0.009	−0.011	0.006	0.013	−0.004
(n = 5,006)	(0.052)	(0.016)	(0.016)	(0.015)	(0.016)	(0.016)	(0.016)
Difference	−0.033	0.038+	0.041+	0.030	0.042+	0.034	0.032
	(0.071)	(0.022)	(0.022)	(0.021)	(0.023)	(0.023)	(0.023)

Table S6. Heterogeneity in treatment effects under Facebook tips by selected covariates. 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 Section 4.1, within specified subgroups. Under two-sided hypothesis tests: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

TK: additional results. @Molly suggests we pin down results in main text first, then add additional material here.