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 16, 2022

Author Contributions:

Competing Interest Statement:

Classification:

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

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.

Abstract:

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 regarding COVID-19-related posts. We found common approaches used by social media platforms, such as flagging misleading posts or including related articles, to be ineffective, and estimate precise null effects of these treatments. Instead, providing tips for spotting misinformation and nudging users to think about the accuracy of media content improves sharing discernment – reducing intended sharing of false posts without adversely affecting true posts. Providing tips leads to effects equivalent to a nearly 8% reduction in intended sharing of false information, and nudging

accuracy to a 4% reduction. Both interventions are successful among some of the worst offenders, who intend to share more false news at baseline. We also find significant differences in response to these treatments across users, indicating differences in mechanisms through which the interventions affect outcomes. The results suggest that these low-cost scalable interventions can significantly improve the quality of information circulating online.

MOW: tried something here

Significance statement: Health misinformation can be deadly. Alongside the global coronavirus pandemic there has been an "infodemic" of rumors about the virus and hoax cures. Which online interventions delivered via social media help to limit the spread of misinformation? We experimentally evaluated a slate of interventions to learn which are effective at reducing social media users' intentions to share false information about COVID-19. Warning labels on posts had little influence while accuracy prompts and tips had positive effects in both countries. The results suggest that platforms may wish to consider implementing these low-cost scalable interventions to reduce sharing of misinformation globally.

1 Introduction

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. Like the actual virus, COVID-19 misinformation is not bounded by state borders. Although other forms of false information have been shown to spread faster and farther than corresponding true information (Vosoughi et al., 2018), early evidence indicates that COVID-19 information from reliable and questionable sources did not differ in their spread (Cinelli et al., 2020). Other work on COVID-19 conspiracy theories, specifically, suggests that these rumors were more viral than neutral or debunking stories (Reis et al., 2020). Regardless of its speed and reach, COVID-19 misinformation presents a challenge for policy makers trying to keep their citizens safe.

Before effective vaccines were developed and widely available people across the globe looked to alternative sources for prevention techniques and remedies for COVID-19, which have proven to be deadly. 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). Particularly in contexts with weak healthcare systems, the uncertainty over how to combat COVID-19 infection may have been particularly daunting and the search for remedies that much more urgent.

This paper focuses on these particularly dangerous pieces of COVID-19 misinformation – hoax "cures" – and tests numerous online interventions designed to curb the spread of these falsities, while not adversely affecting the sharing of true information on COVID-19 prevention techniques. We began this study 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 might naturally engage with similar media posts, to enhance the realism of the study. Participants answered survey questions and were randomized into different treatments delivered by the Messenger chatbot. Our main outcome of interest is sharing discernment – whether respondents indicate wanting to share true but not false posts.

This paper focuses on sharing, rather than belief, since exposure can further false narratives

through resharing even if it doesn't affect an individual's belief. And misinformation can have harmful effects, regardless of the motivation for sharing: there is suggestive evidence that the spread of misinformation is correlated with behavior. One study analyzed more than 100 million Twitter posts worldwide and found a correlation between waves of unreliable information prior to a rise in COVID-19 infections (Gallotti et al., 2020). Another study of Fox News viewers in the US demonstrates that greater exposure to COVID-19 misinformation is associated with lower adoption of preventative behaviors (Bursztyn et al., 2022). Randomized controlled trials also find that exposure to COVID-19 misinformation lowers intentions to engage in physical distancing among a convenience sample of German university students (Pummerer et al., 2022) and intentions get the vaccine among nationally representative samples in the US and UK (Loomba et al., 2021).

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 average social media users. 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 improves beyond convenience samples, laboratory experiments, and opt-in survey panels. But we cannot say how users who decided to participate in our study might 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, or 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). Thus, it is possible that participants may be responding in particular ways due to experimenter demand effects. The validity of our results would be called into question 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. Later we presents tests that suggest our results are not wholly driven by experimenter demand bias.

MOW: not in here yet, currently referenced on p. 14.

Finally, misinformation studies that focus on sharing behavior as the main outcome of interest are inherently limited by ethical concerns of not wanting to contribute to the

ecosystem of misinformation by allowing survey participants to *actually share* false posts. We, like many others, instead use measures of sharing *intentions*. While scholars have found that intentions are correlated with actual online sharing behavior (Mosleh et al., 2020), measuring intentions rather than actual sharing 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.

We should check this is correlated with true sharing intentions.

To curb sharing of false posts we examine interventions delivered to both the individual user, such as tips for spotting fake news and nudges; as well as treatments delivered alongside specific posts, such as flags or warning labels pinned alongside the article of interest. Evaluating both types of interventions – those targeting users and those associated with individual posts – alongside one another is important give how much these different strategies vary in their cost and scalability. Interventions that flag specific posts rely on time and resource intensive fact-checking sources to verify the veracity of individual posts before applying such labels. General interventions delivered to users while they are on a particular platform, on the other hand, are less resource intensive and much more easily delivered en masse.

Traditional randomized experiments are often limited by the number of interventions due to power considerations, but our adaptive design allows us to sort through numerous interventions. Drawing on behavioral science theory and industry practice, we used a multi-factorial adaptive design to evaluate the effectiveness of seven respondent-level interventions and four headline-level interventions during an adaptive learning phase. These interventions speak to the debate as to whether misinformation spreads because people are not paying enough attention or people do not have skills to spot it. We find evidence in support of both theories, and our results suggest that interventions targeting individuals rather than specific posts are more effective overall.

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 in-

terventions 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.

2 Materials and Methods

2.1 Design

Our study was conducted in two stages; a "learning" stage, which ran from Febrary 26 - March 22, 2021, in which we used a multi-factorial contextual adaptive experimental design to learn optimal interventions, and an "evaluation" stage (June 30 - July 20, 2021), in which we compared only the most effective interventions established in the learning stage to obtain precise estimation of their effects.

confirm dates

We considered two types of interventions: seven respondent-level interventions and four headline-level interventions. The respondent-level interventions included behavioral nudges and trainings targeted to the 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 S1 in the Supplementary Information (SI) describes all of the interventions we tested. Section S2.1 in the SI discusses learning stage design and results. The two respondent-level and two headline-level treatments that were found to be most effective in the learning phase include an accuracy nudge and Facebook trips (respondent-level) and factcheck and related articles (headline-level), which are presented in Figure 3.

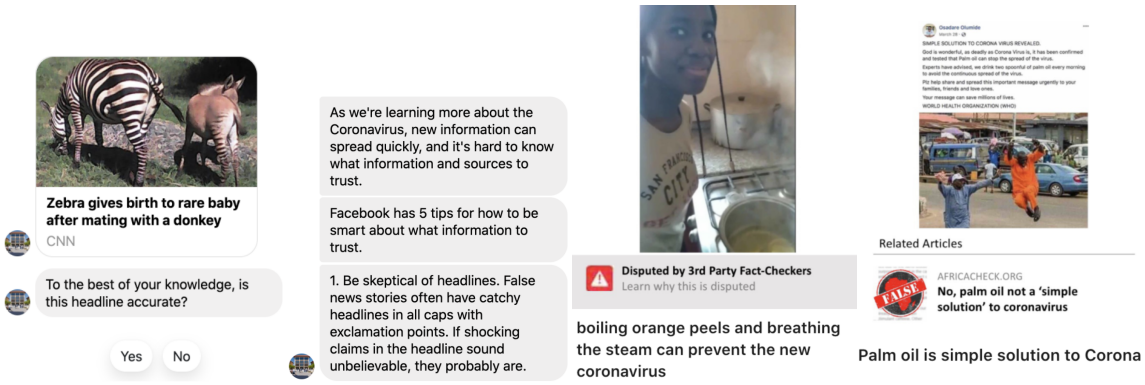

Figure 1. Respondent- and headline-level treatments tested in the evaluation phase.

2.2 Outcome measures

Our outcome measures capture discernment in sharing intentions. We show participants a series of real social media posts about COVID-19 cures, treatments, and preventative best practices and ask if they wanted to share the post. The stimuli include both 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 and was fact-checked in Kenya and Nigeria since the start of the pandemic. Each participants 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.

2.2.1 Combined response measure

In the learning portion of the experiment, our adaptive algorithm updated based on a combined outcome measure, pre-registered in our design document. This measure is the summed number of times users said they would like to share true and misinformation stimuli respectively over Facebook Messenger and on their Facebook Timeline, across two stimuli of each type. Users could share each type of stimuli up to four times (two channels

of sharing x two stimuli). As our aim is to learn treatments that will decrease sharing of false information while not overly harming sharing of true information, false posts are given a weight of -1, and true posts are given a weight of 0.5 in this measure.

LR: helpful to put table of potential values of response funct here? might make later tables/figs easier to interpret?

		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 1. Combined response measure.

2.2.2 Sharing disaggregated by true and false stimuli, and by sharing channel

For our primary reporting, we report results for both types of stimuli separately, for improved interpretability and to better illustrate how the treatments affect sharing discernment. For this measure, we calculate the proportion of true and false stimuli respondents reported intending to share, across either Messenger or timeline. We also report sharing disaggregated by channel (Messenger, timeline).

2.3 Survey recruitment

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. Kenya and Nigeria also represent two of Facebook's top three largest user bashes in sub-Saharan Africa (Africa, 2016), with a combined user base of 30-35 million users ages 18 years and older.[1] We recruited social media users 18 years and over in these countries through targeted Facebook advertisements (see Figure S1 in the SI) (Rosenzweig et al., 2020). Users who clicked on our ads then started a conversation with our page's Messenger chatbot. [2]

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

[2] The ads and experiment were carried out using the Facebook page Social Impact Research Initiative (https://www.facebook.com/socialimpactresearchlab).

2.4 Estimating procedures

For individuals indexed by i, we observe covariates X_i, and use an assignment procedure to assign categorical treatments $W_i \in \mathbf{W}$. Observed response for individual i is represented by Y_i. We denote the outcome for individual i under treatment w as $Y_i(w)$. Treatment assignment probabilities are represented by $e_i(w) := \Pr[W_i = w | X_i = x]$. 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).[3] Covariates used for adjustment are described in further detail in Table S2.

3 Results

3.1 Main effects

Using data from the learning stage (see Supplementary Information), we selected four treatments to test independently in the evaluation stage in addition to the pure control—two

[3]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, which we estimate 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).

respondent-level, *accuracy* and *Facebook tips* and two headline-level, *factcheck* and *related articles* (see Figure). These were the treatments associated with the highest mean responses in each class separate from control, as estimated with the estimator described in Equation 2. We also learned an optimal contextual policy, assigning the most effective of the two respondent-level treatments based on individual covariate profiles. This policy is described in further detail in Section 3.3.

Prior to collecting the evaluation data, we learned a contextual policy based on the combined response function, described in Section 2.2. This optimal contextual policy did not measurably improve the combined response over the two fixed respondent-level policies, and directionally was inferior in decreasing false sharing intentions as compared to Facebook tips. We report outcomes for this policy in Section S2.1 in the supplementary information. To explore the additional benefits of accounting for context with the goal of decreasing false sharing intentions, we report in the main paper results with respect to a policy learned on the false sharing measure only.

3.1.1 Respondent-level treatments

Considering the respondent-level treatments, the accuracy treatment 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 provides 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 Supporting Information.

These treatments relate to two schools of thought as to 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 et al., 2021). A second reasoning suggests that people simply do not know how to identify misinformation. This theory prescribes providing training or

[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). These theories are not the focus of this study, however, and although we did ask about partisanship we do not see substantive differences in treatment effects by party ID.

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.

3.1.2 Headline-level treatments

We also evaluated two headline-level treatments. We used a factcheck intervention that has been used by several platforms and adds "disputed" flags to false posts, and we 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. For COVID-19 information, Kreps and Kriner (2020) find the effectiveness of these tags to depend on context, and worked for only one out of three false COVID headlines they tested. Recent evidence suggests fact checks can improve discernment (belief) in diverse contexts (Porter and Wood, 2021). Brashier et al. (2021) also find that timing 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.

3.1.3 Findings

Table 2 shows 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, we hope to see positive treatment effects for our combined response function, negative treatment effects on false sharing, and positive or neutral treatment effects on true sharing.

		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 2.4. $n = 10,681$. For contrasts only, $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

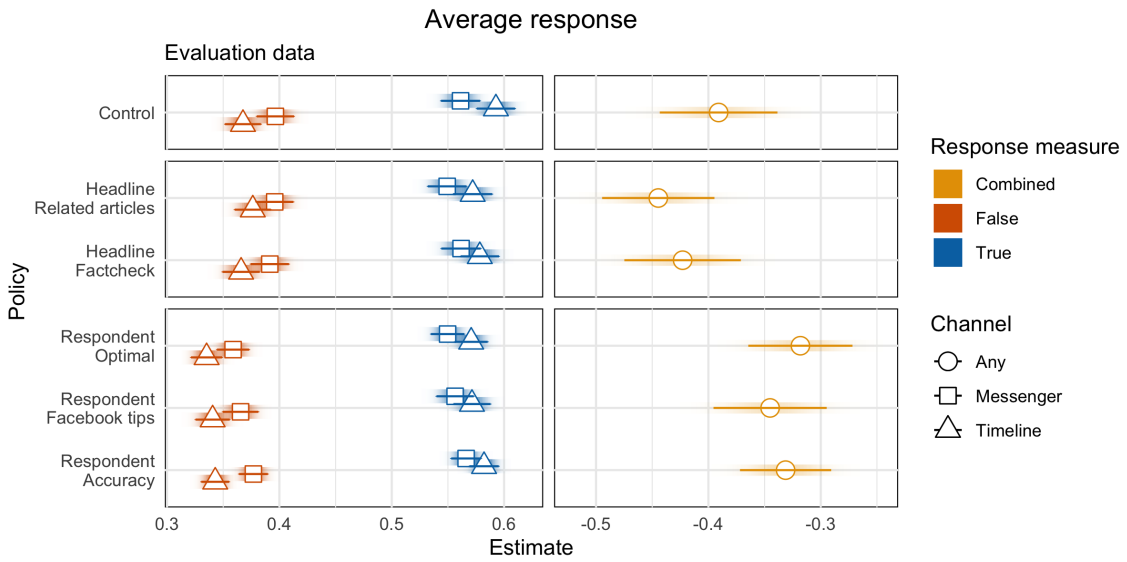

Figure 2. Response estimates. Response measures are average intention to share true and false stimuli over either channel, and a combined response measure, reported in Section 2.2. Estimates are produced from an augmented inverse probability weighted estimator, as described in Section 2.4.

Alternative figure for main results

Under control, we see that respondents exhibit discernment in what types of stimuli they share and over which channels. Respondents report greater intentions to share true stimuli as compared to false stimuli on any channel, and are more likely to share true stimuli on their public timelines as compared to by private message, but the reverse is true for false stimuli. This discernment by channel demonstrates that even when users intend to share false stimuli, they are making different decisions about the way in which they share it.

We pre-specified tests of each type of treatment condition against the control. The two headline-level treatments are not effective towards our objective of 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) 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 more promising. The Facebook tips and accuracy nudge treatments decrease false sharing relative to control by 2.0 pp (SE = 1.0)

and 3.3 pp (SE = 1.1), respectively, with small effects on true sharing, which are not distinguishable from zero at conventional significance levels. While both treatments are effective at decreasing public timeline sharing, the Facebook tips intervention is directionally more effective at also moving private sharing.

Statement of effect size and significance here? Significance on difference in treatment effects across channels is marginal.

Other studies have found similar positive effects of accuracy prompts in diverse settings, 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 be effective 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. Previous research has not differentiated mechanisms driving the Facebook tips treatment, as both treatments appear to be effective for most participants. However, we find evidence of heterogeneity in treatment effects between the two treatments, suggesting that there are some differences in how these treatments work.

To guard against experimenter demand effects, we embedded treatments in a longer survey block about general social media usage. If users' post-treatment response were, however, most responsive based on perceptions of what researchers want, we would expect the headline level treatments to have the largest decreases in false sharing intentions. For these treatments, only false stimuli were accompanied by visual flags. We do not find significant effects here, however. The variation in treatment effects by channel also provides evidence against experimenter demand effects: if users were only responding to perceived experimenter objectives, we would expect effects to be uniform across channels.

3.2 Heterogeneous response and treatment effects

We find covariates are highly predictive of heterogeneity in baseline sharing behaviors. In particular, we focus on several key variables for examining targeting and heterogeneity: 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 of theoretical interest in social scientific research (political allegiance, digital literacy, and scientific knowledge).[5]

[5]Our digital literacy measure is an index of self-reported familiarity with computer and Internet-related terms adapted from Guess and Munger (2020). Trust in science has been found to be a particularly strong

Policymakers with constrained resources may wish to better understand who to target to achieve maximum effect. Hence, it is worth understanding how targeting can be most effectively implemented. One approach to targeting is to identify the greatest culprits of sharing misinformation and direct interventions towards these users. Our data suggest that under control, younger participants, men, participants with low digital literacy, participants aligned with the ruling party, and participants with low scientific knowledge exhibit a higher propensity to share false stimuli (see Table 3).

predictor of belief in COVID misinformation/conspiracy theories (Murphy et al., 2021).

	False			True		
	Any sharing	Messenger	Timeline	Any sharing	Messenger	Timeline
Age						
Below median	0.457	0.413	0.375	0.633	0.547	0.563
(n = 5,412)	(0.012)	(0.012)	(0.011)	(0.012)	(0.012)	(0.012)
Above median	0.425	0.376	0.363	0.671	0.575	0.623
(n = 5,271)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Difference	0.032+	0.038*	0.012	−0.038*	−0.028	−0.060***
	(0.018)	(0.017)	(0.016)	(0.017)	(0.018)	(0.017)
Gender						
Not male	0.400	0.359	0.327	0.611	0.512	0.543
(n = 5,050)	(0.012)	(0.012)	(0.011)	(0.012)	(0.013)	(0.012)
Male	0.479	0.426	0.407	0.687	0.604	0.638
(n = 5,633)	(0.012)	(0.012)	(0.012)	(0.011)	(0.012)	(0.012)
Difference	−0.078***	−0.078***	−0.078***	−0.076***	−0.076***	−0.076***
	(0.018)	(0.018)	(0.018)	(0.017)	(0.017)	(0.017)
Political allegiance						
Not aligned	0.416	0.365	0.341	0.630	0.535	0.564
(n = 7,570)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
Aligned (n = 3113)	0.505	0.467	0.438	0.704	0.622	0.662
	(0.016)	(0.016)	(0.015)	(0.015)	(0.016)	(0.016)
Difference	−0.090***	−0.090***	−0.090***	−0.074***	−0.074***	−0.074***
	(0.019)	(0.019)	(0.019)	(0.018)	(0.018)	(0.018)
Digital literacy index						
Below median	0.495	0.448	0.423	0.675	0.587	0.621
(n = 5,443)	(0.012)	(0.012)	(0.012)	(0.011)	(0.012)	(0.012)
Above median	0.387	0.340	0.314	0.627	0.533	0.564
(n = 5,240)	(0.012)	(0.012)	(0.011)	(0.012)	(0.013)	(0.013)
Difference	0.108***	0.108***	0.109***	0.048**	0.055**	0.058***
	(0.018)	(0.017)	(0.016)	(0.017)	(0.018)	(0.017)
Scientific knowledge index						
Below median	0.459	0.412	0.384	0.659	0.561	0.597
(n = 5,677)	(0.012)	(0.012)	(0.011)	(0.012)	(0.012)	(0.012)
Above median	0.423	0.375	0.352	0.643	0.560	0.588
(n = 5,006)	(0.013)	(0.012)	(0.012)	(0.012)	(0.013)	(0.013)
Difference	0.036*	0.036*	0.031+	0.016	0.002	0.010
	(0.018)	(0.017)	(0.016)	(0.017)	(0.018)	(0.017)

Table 3. 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 2.4, within specified subgroups. For differences only: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

For these groups we find that assigning either the accuracy nudge or the Facebook tips treatment 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 only statistically significant when comparing users with low to those with high levels of scientific knowledge.

[TK: comments about heterogeneity in channel sharing.]

	False			True		
	Any sharing	Messenger	Timeline	Any sharing	Messenger	Timeline
Age						
Below median	−0.018	−0.021	−0.018	−0.025+	−0.019	−0.035*
(n = 5,412)	(0.014)	(0.013)	(0.012)	(0.014)	(0.014)	(0.014)
Above median	−0.035*	−0.026+	−0.038**	0.009	0.019	0.002
(n = 5,271)	(0.014)	(0.013)	(0.013)	(0.013)	(0.014)	(0.014)
	(0.014)	(0.013)	(0.013)	(0.013)	(0.014)	(0.014)
Difference	0.016	0.005	0.021	−0.034+	−0.038+	−0.037+
	(0.019)	(0.018)	(0.018)	(0.019)	(0.019)	(0.019)
Gender						
Not male	−0.019	−0.022+	−0.021+	−0.024+	−0.004	−0.029*
(n = 5,050)	(0.014)	(0.013)	(0.012)	(0.014)	(0.014)	(0.014)
Male	−0.033*	−0.024+	−0.034**	0.006	0.004	−0.005
(n = 5,633)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
Difference	0.014	0.002	0.013	−0.030	−0.007	−0.025
	(0.019)	(0.018)	(0.018)	(0.019)	(0.019)	(0.019)
Political allegiance						
Not aligned	−0.035**	−0.029**	−0.040***	−0.007	0.004	−0.015
(n = 7,570)	(0.011)	(0.011)	(0.011)	(0.011)	(0.012)	(0.012)
Aligned	−0.006	−0.009	0.001	−0.012	−0.009	−0.019
(n = 3,113)	(0.018)	(0.017)	(0.017)	(0.017)	(0.018)	(0.017)
Difference	−0.029	−0.020	−0.041*	0.005	0.012	0.004
	(0.021)	(0.020)	(0.020)	(0.021)	(0.021)	(0.021)
Digital literacy index						
Below median	−0.043**	−0.039**	−0.038**	−0.025+	−0.012	−0.025+
(n = 5,443)	(0.014)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
Above median	−0.009	−0.007	−0.018	0.008	0.012	−0.007
(n = 5,240)	(0.014)	(0.013)	(0.012)	(0.014)	(0.014)	(0.014)
Difference	−0.033+	−0.032+	−0.020	−0.033+	−0.024	−0.018
	(0.019)	(0.018)	(0.018)	(0.019)	(0.019)	(0.019)
Scientific knowledge index						
Below median	−0.045***	−0.044***	−0.040**	−0.026*	−0.011	−0.033*
(n = 5,677)	(0.013)	(0.013)	(0.012)	(0.013)	(0.014)	(0.013)
Above median	−0.006	0.001	−0.014	0.012	0.013	0.002
(n = 5,006)	(0.014)	(0.013)	(0.013)	(0.013)	(0.014)	(0.014)
Difference	−0.039*	−0.045*	−0.026	−0.038*	−0.024	−0.035+
	(0.019)	(0.018)	(0.018)	(0.019)	(0.019)	(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 2.4, within specified subgroups. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

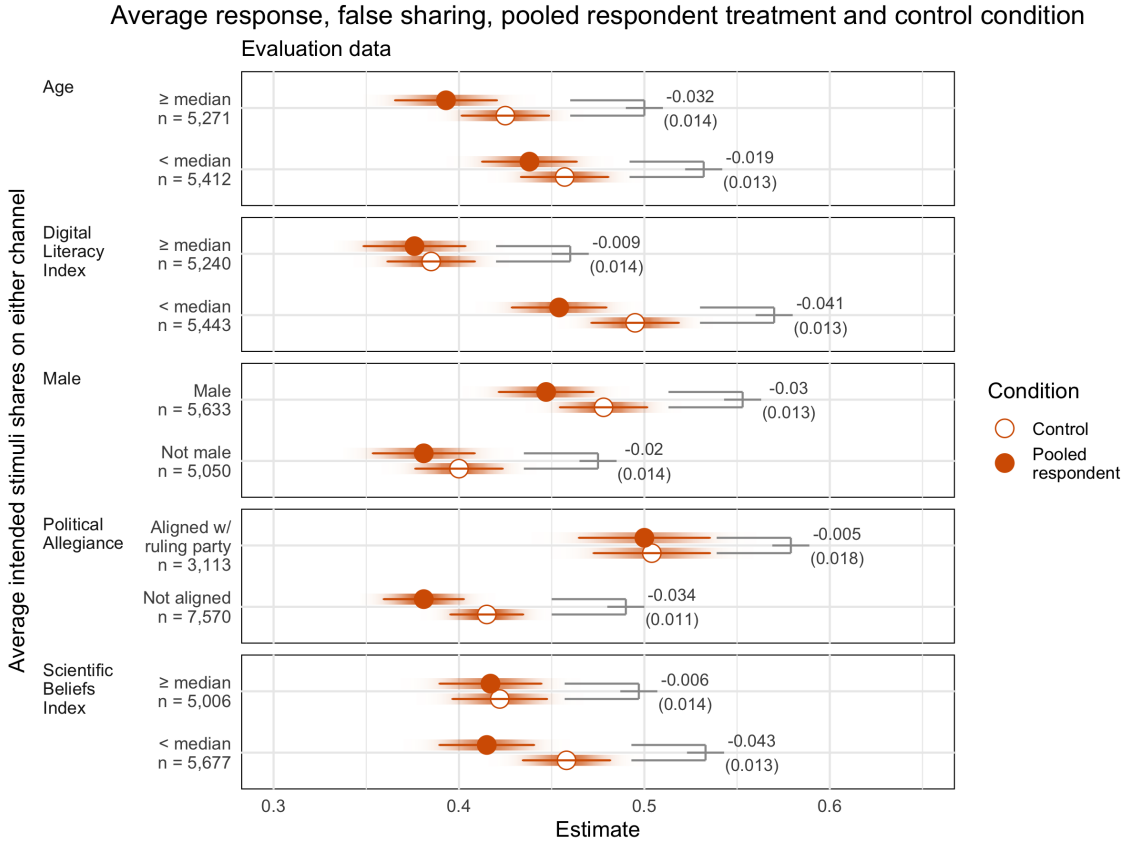

Figure 3. Response under the control condition. The outcome measure is the proportion of true or false stimuli participants reported wanting to share, either as a Facebook post or privately in Facebook Messenger. Estimates are produced from an augmented inverse probability weighted estimator, as described in Section 2.4, within specified subgroups.

Note that this table represents **treatment effect estimates** in comparison to control, whereas Table 2 represents control mean. Is this confusing? Better way to differentiate?

replicate table with treatment effects as percent change over baseline

Alternative figure

3.3 Heterogeneity in best policy

We can also consider targeting in terms of which participants should get which treatment. We estimate average decreases in false sharing intentions for both the accuracy nudge and the Facebook tips treatment. However, we also observe differences in how users respond to these two treatments. Figure 4 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 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 4, is -3.8 pp (SE = 1.3), using the targeting operator characteristic curve.

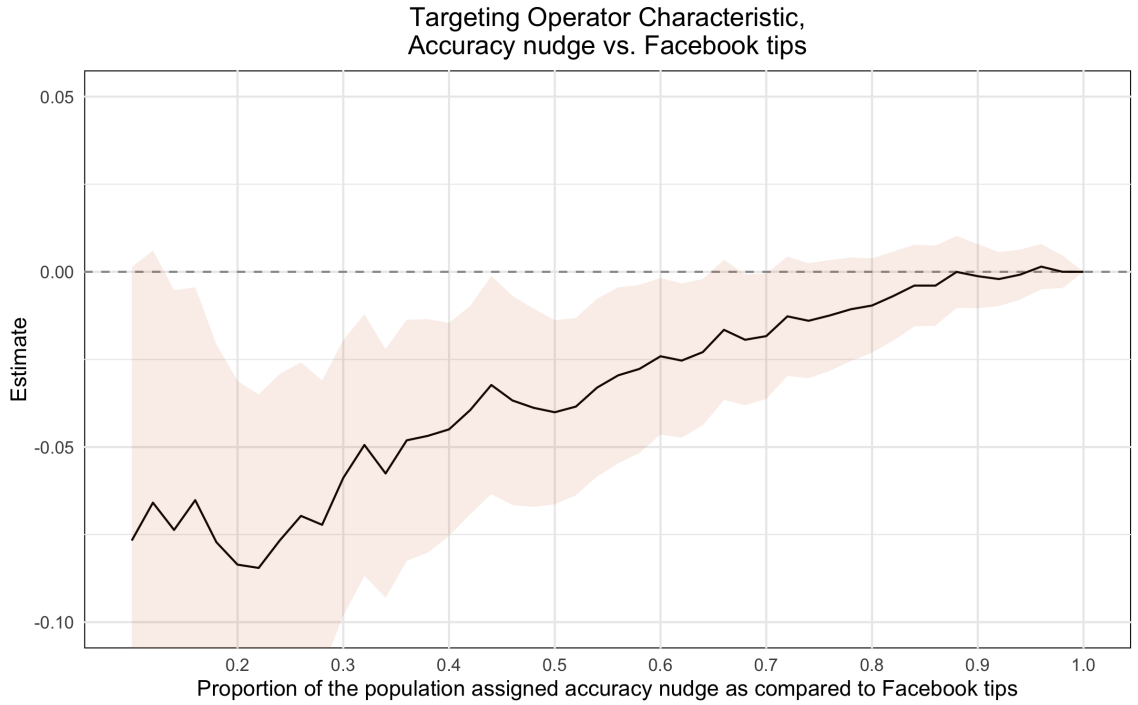

Figure 4. Targeting operator characteristic curve, comparing the accuracy nudge and Facebook tips. The outcome measure is the difference in proportion of false stimuli participants reported wanting to share, either as a Facebook post or privately in Facebook Messenger, between the accuracy nudge and 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 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.

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. 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 (see Figure 5).

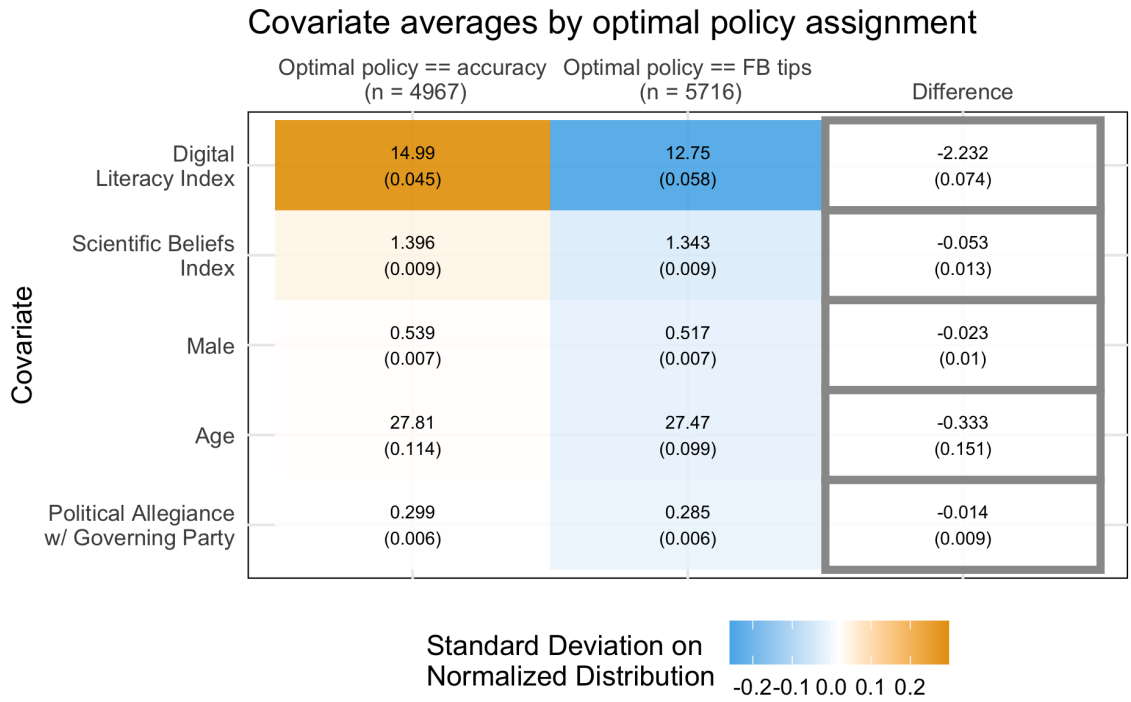

Figure 5. Selected covariate means between participants assigned to the accuracy nudge as compared to participants assigned to Facebook tips under the contextual policy. Covariates are ordered by size of standardized deviation between the two groups.

Should this figure just be a table?

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 (one-sided p-values for difference of 0.002 and 0.130, respectively).

In Table 5, 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 -1.7, SE = 1.5). The heterogeneity in treatment effects between the two groups (difference of 3.8, SE = 1.4) provides new evidence that participants respond differently to these two treatments.

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

Table 5. 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 2.4, within specified subgroups. For differences only: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4 Discussion

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. Such policies delivered to participants are much more cost effective than headline-specific interventions that require time and effort from human or AI fact checkers.

Ours is one of only a few studies examining Facebook's related articles policy (see also Bode and Vraga, 2015), hence more research is needed. The fact that we do not see effects from fact-checking interventions in this setting may raise concerns given the existing evidence of disputed labels and warnings from the US and other high-income countries. However, we believe that these results may vary across settings given the variations in context, specifically the general levels of digital literacy and abilities to discern fact from fiction. For instance, when we analyze the headline-level interventions by baseline-levels of sharing "discernment" (pre-treatment sharing of true $>$ false information), we find that there is a group of participants for whom showing related articles under false posts reduces false sharing compared to control.

Specifically, participants with pretreatment sharing discernment above the sample median share false news less under related articles treatment as compared to control ($p < 0.05$), participants below median share false news more under related articles ($p < 0.06$), and the difference is significant at $p < 0.01$. Participants with above median baseline levels of sharing discernment share false stimuli (along either channel) 19.1 percent of the time. Under the related articles treatment, this sharing is decreased by 1.4 pp, a 7.2 percent reduction. These findings perhaps hint at the idea that more ambiguous post-level interventions such as showing related articles are only effective among already well-discerning individuals. This may also help to explain our general finding of no effects from headline-level treatments among our sample—who may have lower baseline discernment than samples from high-income countries that may have generally higher levels of education or have greater digital literacy. More research from diverse contexts

is required to understand which interventions are effective and why some treatments that work in certain settings are not effective in others.

5 Acknowledgements

[TK: Facebook Health; James, Ricardo, Undral; feedback from various workshops.]

References

Africa, I. (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.

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.

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.

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*.

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., and Scala, A. (2020). The covid-19 social media infodemic. *Scientific reports*, 10(1):1–10.

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.

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.

Gallotti, R., Valle, F., Castaldo, N., Sacco, P., and De Domenico, M. (2020). Assessing the risks of 'infodemics' in response to covid-19 epidemics. *Nature human behaviour*, 4(12):1285–1293.

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.

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.

Guess, A. M. and Munger, K. (2020). Digital literacy and online political behavior. *Political Science Research and Methods*, pages 1–19.

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

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

Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., and Larson, H. J. (2021). Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature human behaviour*, 5(3):337–348.

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. (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.

Pummerer, L., Böhm, R., Lilleholt, L., Winter, K., Zettler, I., and Sassenberg, K. (2022). Conspiracy theories and their societal effects during the covid-19 pandemic. *Social Psychological and Personality Science*, 13(1):49–59.

Reis, J. C. S., Melo, P., Garimella, K., and Benevenuto, F. (2020). Can whatsapp benefit from debunked fact-checked stories to reduce misinformation? *The Harvard Kennedy School (HKS) Misinformation Review*.

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

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

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.

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

We conduct this study with social media users in two major English-language hubs of online communication in sub-Saharan Africa, Kenya and Nigeria. Facebook estimates that there are 30-35 million Facebook users who are 18 years and older from these two countries (as reported on the audience insights tool on Facebook's advertising platform). AfricaCheck.org, a third party verification site, has offices in both countries and has recently created pages devoted to coronavirus-related misinformation circulating online. From January to March, the number of English-language "fact-checks" (i.e., publicly spread pieces of information deemed false or misleading by fact-checking organizations) increased by more than 900% worldwide (Brennen et al.,2020), demonstrating the prevalence of this kind of content and the availability of verified COVID-related information. Both countries also have at some point had access to Facebook zero, a light free version of Facebook, meaning that users do not need to pay for data to access certain features of the platform.

As is becoming increasingly popular among social scientists, we recruited social media users 18 years and over in Kenya and Nigeria through targeted Facebook advertisements (Rosenzweig et al., 2020). 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. In contrast to sending users to an external survey platform such as Qualtrics, the benefit of the chatbot is that we keep users on the Facebook platform, with which they are likely more familiar, and maintain a realistic setting in which users might encounter online misinformation. Participants who completed the survey in the chatbot received compensation in the form of mobile phone airtime (equivalent to about $0.50) sent to their phone.

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), 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 (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 (see **??**).
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 S1. 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 S2. Covariates and response options

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?

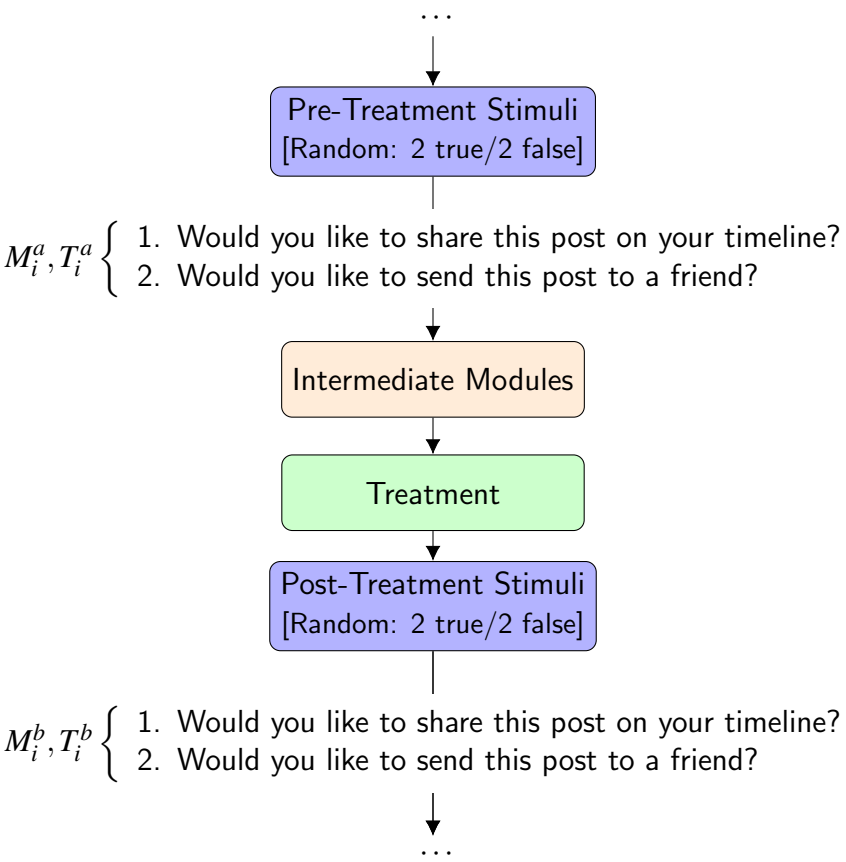

Figure S2. Survey flow.

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.

S2 Additional results

S2.1 Learning stage

Revise wording for this section based on Susan's feedback

In the learning stage of our study, we treated respondent-level and headline-level treatments as separate treatment factors, each with a baseline control condition, to facilitate learning about treatment interactions. We used a contextual adaptive algorithm that updated treatment assignment probabilities dynamically during the experiment, assigning treatment to

each individual based on their predicted response under alternative treatment conditions, conditional on covariates (see Figure S3).

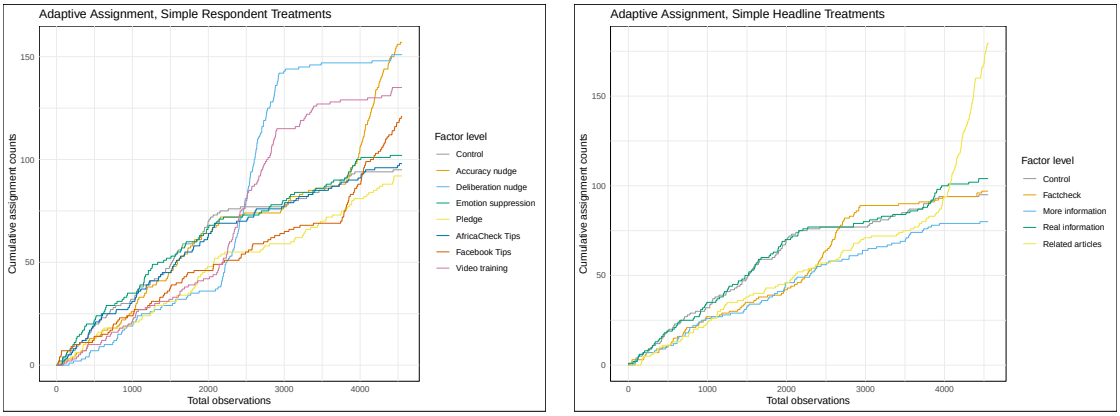

Figure S3. Cumulative treatment assignment during the learning phase for headline (left panel) and respondent (right panel) interventions. While the full design allows for all factor combinations, these plots illustrate assignment using the "pure" version of each factor, i.e., when the other factor is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

show something about the bayesian posteriors too?

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.

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