News credibility labels improve news diets, reduce misperceptions, and increase media trust*

Kevin Aslett[a], Andrew M. Guess[b], Jonathan Nagler[a,c], Richard Bonneau[a,d], and Joshua A. Tucker[a,c]

[a]Center for Social Media and Politics, New York University
[b]Department of Politics, Princeton University
[c]Wilf Family Department of Politics, New York University
[d]Department of Biology, New York University

Abstract

As the primary arena for viral misinformation shifts toward transnational threats such as the Covid-19 pandemic, the search continues for scalable, lasting countermeasures compatible with principles of transparency and free expression. To inform future interventions, we conducted a randomized controlled trial of simple source credibility labels embedded in users' social feeds and search results pages. Combining surveys and digital trace data from a subset of respondents, we find that after three weeks of embedded source contextual information, real-world exposure to low-quality news sources decreased by nearly 36%, and the perceived accuracy of popular misinformation spread about the Black Lives Matter movement and Covid-19 also measurably declined. Unlike some other strategies to counteract misperceptions, our intervention seemed to achieve these effects without reducing the perceived accuracy of true information. Finally, the intervention led to increased trust in the mainstream media, raising the possibility of a self-sustaining cycle of improved news diets and more positive attitudes toward credible news outlets.

*We are extremely grateful to Craig Newmark Philanthropies for supporting this research project. The Center for Social Media and Politics at New York University is generously supported by funding from the National Science Foundation, the John S. and James L. Knight Foundation, the Charles Koch Foundation, the Hewlett Foundation, Craig Newmark Philanthropies, the Siegel Family Endowment, and NYU's Office of the Provost and Global Institute for Advanced Study. NewsGuard did not consult with the authors on the study design or provide funding support for this research. This study has been approved by the Princeton University Institutional Review Board (#12800). Special thanks to Sam Luks at YouGov.

The internet and social media have drastically decreased the cost of disseminating information by reducing reliance on traditional gatekeepers. As a consequence of this openness and availability, news and information sources have flourished from a variety of ideological and cultural perspectives. The resulting cacophony has encouraged participation by previously underrepresented voices and enabled criticism of dominant authorities. At the same time, it has intersected with existing political divides in ways that have contributed to pathologies in American political discourse including the spread of misinformation (Lazer et al. 2018; Vosoughi et al. 2018; Grinberg et al. 2019; Guess et al. 2019), disagreements about basic facts related to governance and policy (Flynn et al. 2017; Anspach et al. 2019; Pennycook and Rand 2021), and lowered trust in established media (Guess et al. 2021). Of particular concern is the possibility that these problems are interlinked: As political divisions widen, partisan media alienate people from authoritative sources, which could make it more difficult to counteract potentially corrosive — and in the case of public health during a pandemic, life-threatening (Brennen et al. 2020) — misinformation.

Over the past several years, scholars, technologists and policy makers have proposed a number of solutions intended to reduce exposure to misleading information. These range from relatively intrusive measures such as algorithmic downranking, to subtle warnings and labels targeted at specific factual claims (Ecker et al. 2010; Clayton et al. 2019), to general efforts to boost digital media literacy skills (Guess et al. 2020b). A key challenge in these efforts is how to balance the strength of an intervention with potential negative externalities in the form of unintended spillover effects (Pennycook et al. 2020; Nyhan et al. 2013)or limits on individual autonomy and freedom of expression. With this tension in mind, we focus on dynamic feedback in the form of informational labels designed to educate people about the quality of sources that they consume and view in their search or social media feeds (Lorenz-Spreen et al. 2020). This approach builds on humans' tendency to rely on cognitive shortcuts and heuristics, which depending on context can be relatively informative (in the case of source transparency; see Iyengar and Hahn 2009; Gigerenzer and Selten 2002; Pennycook and Rand 2019a) or potentially distorting (in the case of social cues, which are a common feature of social media; see Messing and Westwood 2014).

Aside from providing critical policy-relevant evidence, the efficacy of this type of interactive feedback sheds light on the determinants of people's information diets. One set of arguments contends that people's online news consumption is largely determined by their political preferences — a perspective that, when translated to the domain of politics, implies that many people reside in relatively impermeable, non-overlapping informational bubbles (Sunstein 2017). In this view, political judgments as well as factual beliefs could be powerfully shaped by strong and persistent differences in partisans' news diets. A revisionist literature has emerged challenging many of these claims, which suggests that most people (with the exception of some strong partisans) regularly encounter cross-cutting information in their online browsing activity and

social media feeds, and consumption habits are less ingrained than they first appeared. (Gentzkow and Shapiro 2011; Bakshy et al. 2015; Eady et al. 2020; Fletcher et al. 2020; Guess 2021). If news consumers are receptive to informational nudges, this suggests that media consumption habits are driven at least in part by environmental factors, such as online choice architecture, that can be altered (Thaler and Sunstein 2009). This study stands in a unique position to contribute to this debate considering that no empirical work, to our knowledge, has demonstrated that an intervention designed to help reduce misinformation, in our case contextual "nudges," actually shifts online news consumption patterns.

Prior research suggests mixed expectations about whether subtle, embedded source information cues can change online news consumption behavior or reduce misperceptions. As suggested, there is limited direct evidence that relatively light-touch interventions designed to pre-bunk or directly challenge misinformation can affect downstream online behavior (though for an exception related to link sharing on Twitter, see Pennycook et al. 2021). At the same time, behavioral-science approaches suggest potential benefits of quality heuristics in limited-information environments (Lupia 1994; Gigerenzer et al. 2011; Lorenz-Spreen et al. 2020). Looking beyond potential news diet changes, whether source-level information can inform people's judgments about the veracity of claims is also uncertain given existing evidence on the effectiveness of general warnings about "fake news" (Clayton et al. 2019). But even if perceived accuracy of false claims is reduced as a result of our intervention, this effect might be accompanied by a corresponding decrease in perceived accuracy of true claims as well, as studies have sometimes found in the case of digital media literacy tips (Guess et al. 2020b).

To study this approach, we build on recent innovations for rigorously evaluating online tools (Munzert et al. 2021a,b). We design a pre-registered field experiment that randomly encourages participants to install a prominent web browser extension, NewsGuard, that embeds straightforward source-level indicators of news reliability into users' search engine results pages (SERPs), social feeds and visited URLs.[1] Different "shield" symbols are placed in-feed to provide visual summaries of sources' quality. A green shield indicates a reliable source (examples include CNN, Fox News, and *The Washington Post*), a red shield indicates an unreliable source (examples include Gateway Pundit, *Epoch News*, and Daily Kos), a gray shield indicates a source with user-generated content (such as YouTube, Wikipedia, and Reddit), and a gold shield represents satire (such as *The Onion*, *Babylon Bee*, and *The Daily Mash*). The user can click on the shield to receive more detailed information about the reliability of the news domain in question.[2]

[1] NewsGuard launched in 2018 and produces ratings based on neutral criteria evaluated by a team of journalists and editors; more information can be found in the Materials and Methods section of this paper and at `www.newsguardtech.com`. Although this study employed the NewsGuard extension, which was freely available in app stores at the time of fielding, NewsGuard did not provide any financial support or assistance in the design of our study.

[2] We show these source reliability symbols in Section G of the Supplementary Materials.

We ran a pre-registered[3] field experiment, drawing on a representative online sample of Americans, in which we encouraged a subset of survey respondents to install the NewsGuard web extension. Our main hypotheses test whether in-feed source reliability labels shift downstream news and information consumption from unreliable sources known for publishing misleading or false content to more reliable sources (H1), increase trust in mainstream media and reliable sources (H2), and mitigate phenomena associated with democratic dysfunction (affective polarization and political cynicism) (H3).

We also consider three research questions for which our *a priori* expectations were less clear. First, past research suggests that certain kinds of interventions can reduce people's beliefs in both accurate and inaccurate information (Clayton et al. 2019; Guess et al. 2020b), so we examine whether respondents encouraged to install the NewsGuard extension were more or less likely to believe popular false and true stories that spread during the treatment period.[4] Second, we explore whether downstream effects occur on other outcomes such as trust in institutions, belief that "fake news" is a problem in general, and belief that "fake news" is a problem in the mainstream media. Third, we explore whether any of the identified effects are greater among subgroups found in prior research to more frequently engage with online misinformation.[5] Results from all of our pre-registered analyses can be found in the Supplementary Materials, Sections C and D.

Using panel survey data, behavioral measures of treatment compliance, and individual-level web visit data, we find that in-browser contextual source cues (1) shift participants' online consumption from unreliable sources known for publishing misleading or false content to more reliable sources on average; (2) reduce average belief in widely circulated inaccurate claims without reducing belief in accurate claims; and (3) increase trust in the media generally. These findings suggest that embedding informational feedback about news publishers in users' browser sessions can nudge them away from publishers of misinformation and help them distinguish between false and true news content, results with important implications for scientific research on the efficacy of informational cues, the causes and consequences of information diets, and, more broadly, for the development of neutral, evidence-based interventions to improve the quality of people's news diets.

Results

To measure the effect of in-feed source labels, we conducted a two-wave panel survey in summer 2020 (Wave 1: May 28–June 9, $N = 3,862$; Wave 2: June 19–June 30, $N = 3,337$) that included a randomized

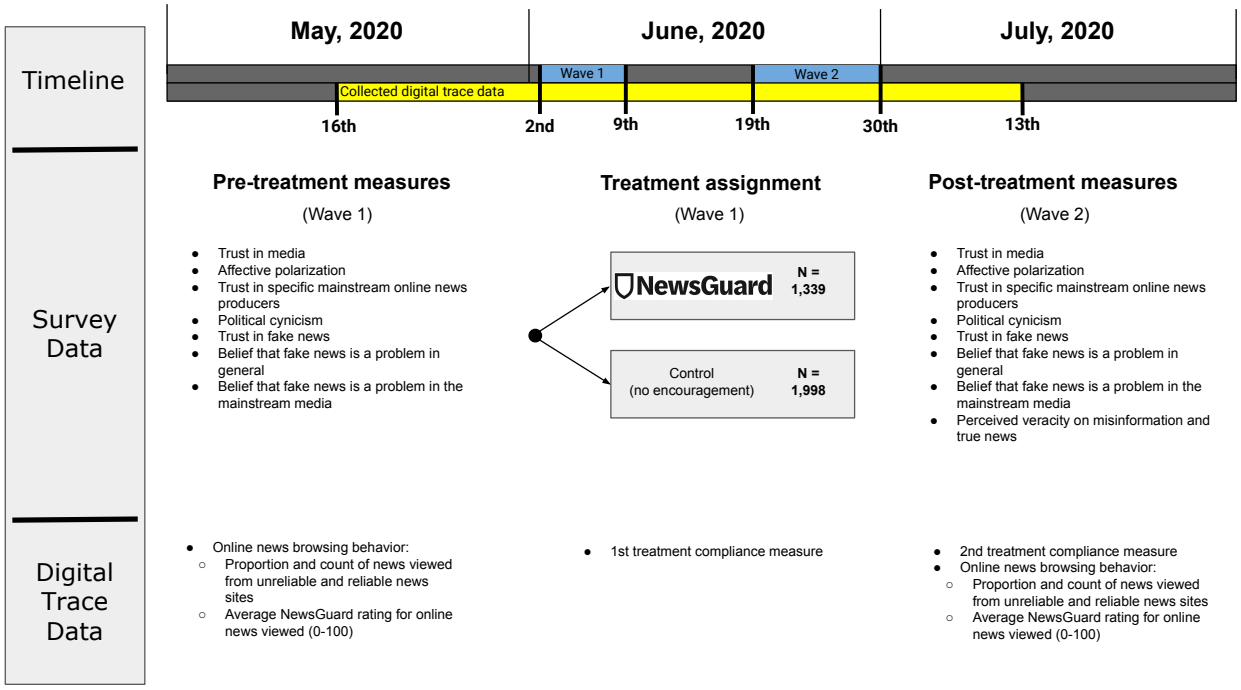
[3]The pre-registration can be found here: `https://osf.io/9qrkt/`

[4]This research question was not pre-registered because we selected the items as close to fielding as possible. We include the results because they are more directly comparable to studies evaluating the effects of interventions designed specifically to counteract misperceptions.

[5]These groups include those who use social media sites, have low levels of digital literacy, consume more news, and already visit more online publishers of untrustworthy news.

incentive to install the NewsGuard web extension at the beginning of the first wave. Fig. 1 presents an overview of the study design. In addition to studying survey-based outcomes, we analyze linked digital trace data to measure the quality of news consumption of a subset of our participants. We create five distinct measures of news diet quality[6] over three time periods: (i) the period before a respondent was assigned treatment in the Wave 1 survey; (ii) the period from treatment assignment to June 30; (iii) the nearly two-week period from July 1 to July 13. We leverage the exogenous disabling of NewsGuard's free capabilities on July 1[7] to determine whether the behavioral effects of this intervention decay after its features are no longer available (Gerber et al. 2011), or if the intervention has more durable effects like other novel informational nudges (Coppock et al. 2018).

Figure 1: Overview of Study Design.



In this section we primarily report covariate-adjusted estimates of Complier Average Causal Effects (CACE) that instrument receipt of the NewsGuard treatment with an indicator for the treatment assignment. We measured whether participants in the treatment and control groups had installed and activated the NewsGuard extension on their web browsers twice: directly after the treatment was assigned in Wave 1 and in the last week of the treatment period.[8] We consider respondents who are found to have the extension

[6] We calculated the average NewsGuard reliability score for websites visited, proportion and counts of unreliable (NewsGuard score < 60) news sites visited, and proportion and counts of reliable news sites visited.

[7] The NewsGuard extension became a pay service with a monthly subscription fee of $2.99.

[8] More details on how compliance was estimated, compliance rates, and how compliers differed from non-compliers can be found in the Materials and Methods section. About 1% of those in the control group already had the NewsGuard extension installed and activated.

running both times to be "treated." The CACE estimand, as compared to Intent-to-Treat (ITT) effects, more directly speaks to a counterfactual policy regime in which a web browser or social platform incorporated credibility labels as part of the user experience.[9]

Consistent with our first hypothesis (H1), randomized exposure to in-browser source reliability information shifts online consumption of news away from unreliable publishers. The treatment effect estimates for each pre-registered behavioral outcome among those who installed the browser extension as a result of the random encouragement are presented in Figure 2.[10] As Figure 2 indicates, we estimate a 1-percentage-point ($\beta = -0.0117$, SE $= 0.0044$; $P < 0.01$) decrease in the proportion of news consumed from unreliable sources in the two weeks after NewsGuard was disabled. Since the average proportion of unreliable news consumed by respondents in the pre-treatment period was 0.03, the effect of source reliability labels represents a nearly 36% decrease in the proportion of unreliable news consumed. The effect is weaker when measured between treatment assignment and July 1, when the extension's functionality ended, and not statistically distinguishable from zero, which is consistent with an over-time learning effect as respondents receive more feedback about the reliability of online sources in their search results and social feeds.[11] In addition to a reduction in the proportion of unreliable news consumed, we also find an approximately 1-point ($\beta = 1.11$, SE $= 0.497$; P < 0.05) increase in the average NewsGuard reliability score of news consumed beginning July 1 relative to the control group. Given that the average reliability score of online news consumed by all respondents from whom we collected digital trace data was 87.6, this represents a 1.3% increase. A similar effect is found in the period between treatment assignment and July 1. Our other behavioral measures did not measurably change in either time period due to the treatment, indicating that the intervention was more effective at shifting online news consumption away from unreliable news than increasing consumption of "reliable" news.

The strongest reduction in the consumption of unreliable online news was among those who already consumed more unreliable news prior to treatment assignment. Figures 3a and 3b present scatter plots with a linear regression line for the proportion of unreliable news consumed (Figure 3a) and average reliability scores (Figure 3b) in the period before the treatment was assigned (x-axis) and between July 1–13 for each respondent. The regression lines suggest that the effect was not consistent across all groups. Rather, those who consumed the most unreliable news (as measured by the proportion of unreliable news and the average reliability score of news consumed) were most likely to consume less unreliable news from July 1–13 as a result of the intervention, relative to the pre-treatment period.

[9] We report estimates of the ITT and CACE in both covariate-adjusted and unadjusted models, as well as CACE estimates using a more lenient standard for treatment uptake, in the Supplementary Materials.

[10] Throughout the discussion, these correspond to CACE estimates using the strictest measure of treatment uptake; see Materials and Methods. See the Supplementary Materials and Methods sections C1 and C2 and sections D1 and D2 for full model results and results from the ITT model and the CACE model using a looser definition of treatment uptake (both covariate-adjusted and unadjusted models).

[11] A figure depicting this is presented in the Supplementary Materials and Methods in Section F.

Figure 2: This figure presents estimates of the effect of the intervention (with 95% confidence intervals) on our pre-registered online behavioral measures in the two periods after treatment assignment: before July 1 when the NewsGuard extension was available and the two-week period between July 1–13 when the NewsGuard extension was disabled. The effect is reported in standard deviations of that measure (pre-treatment).

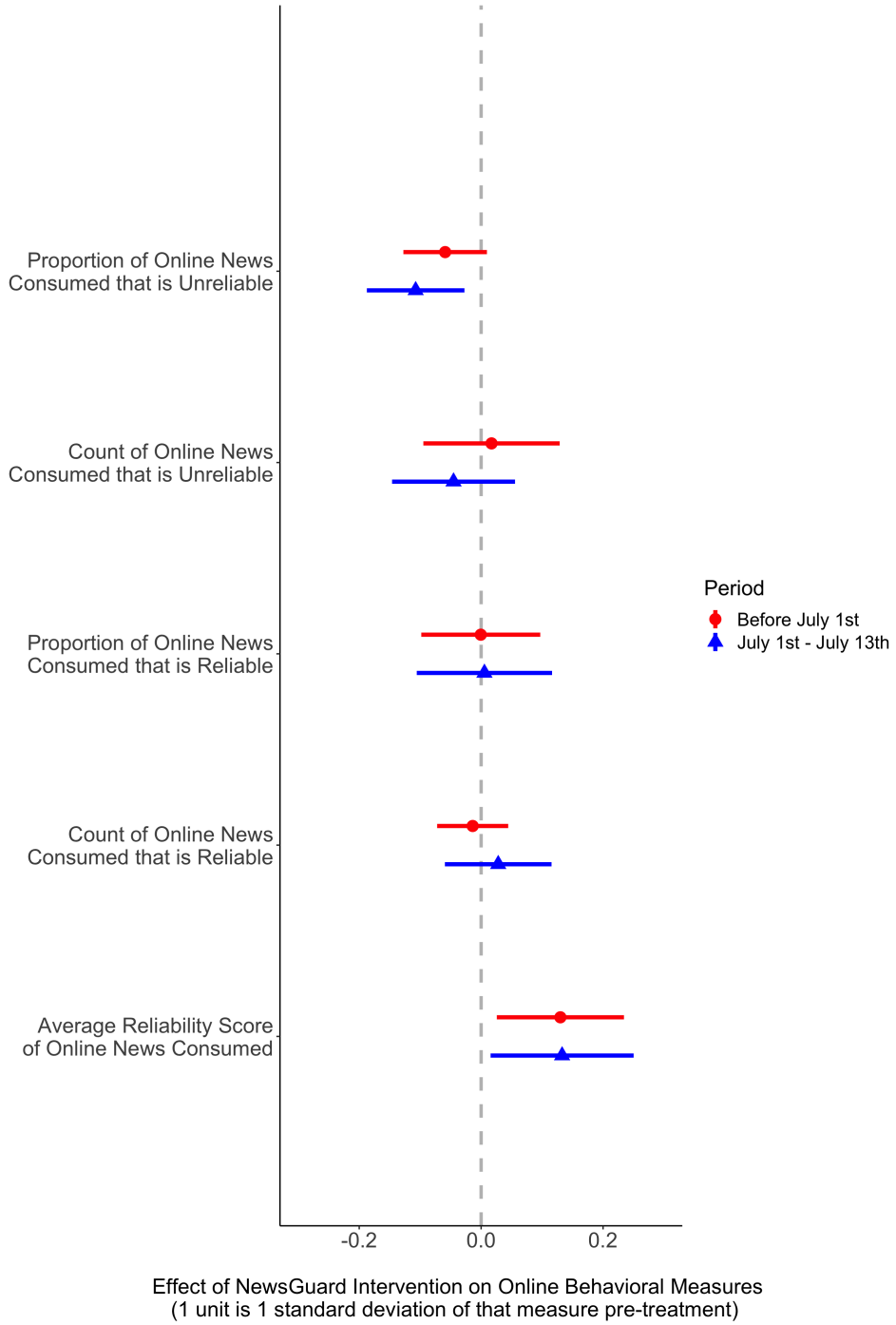
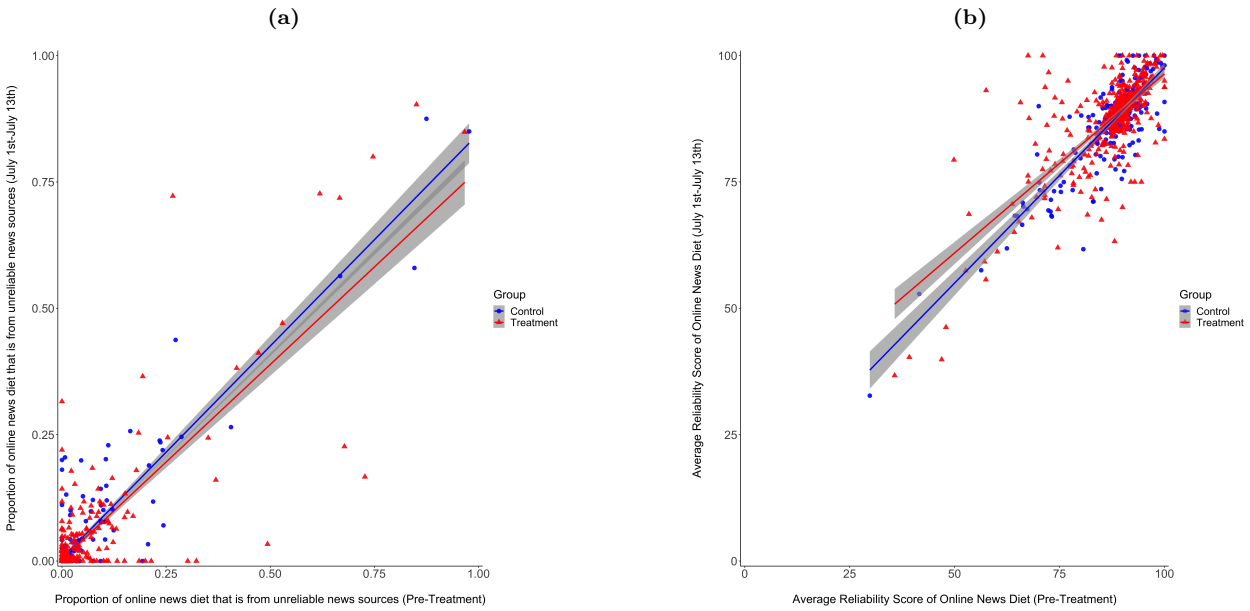


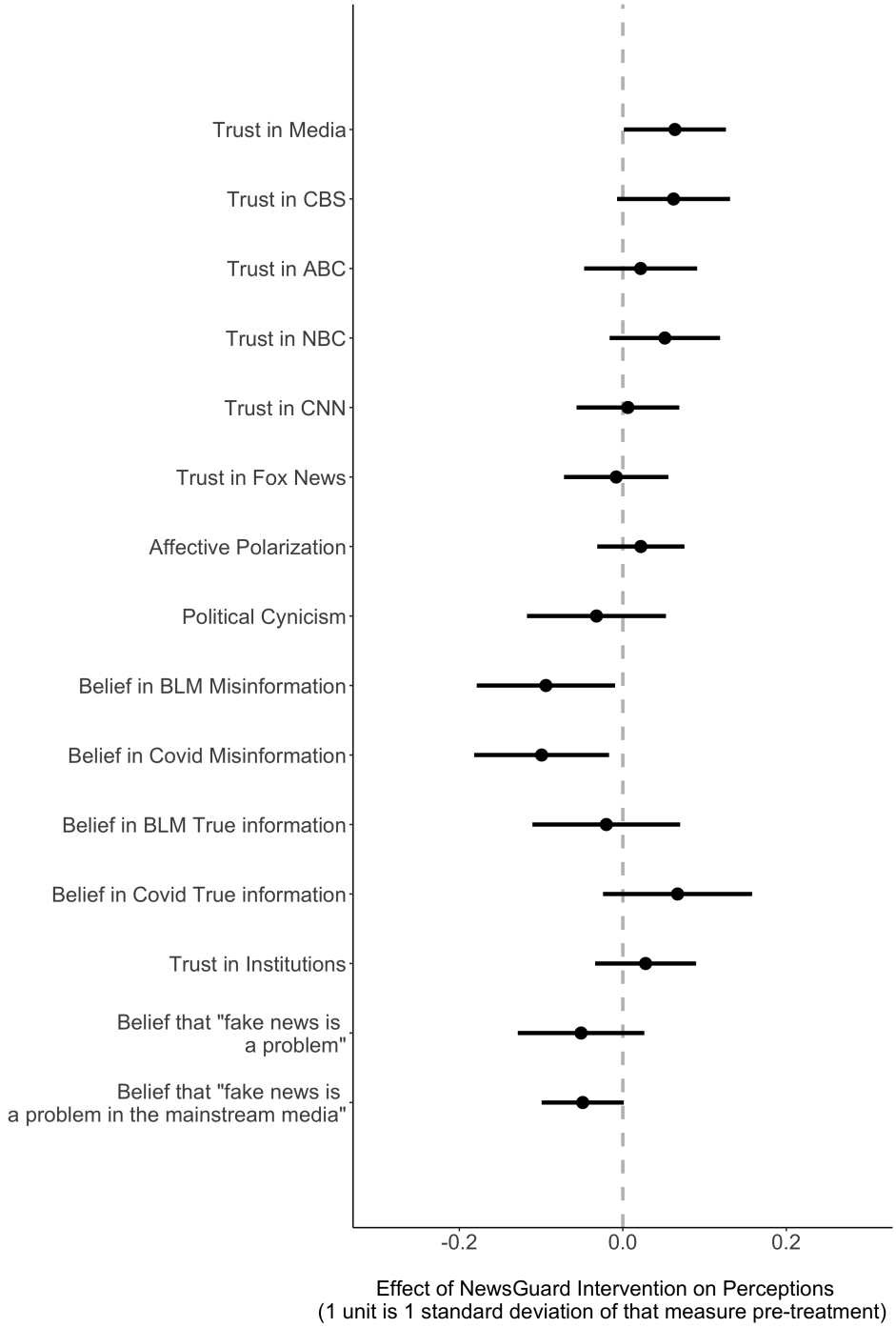
Figure 3: These figures plots the the proportion of news diet that is unreliable (Figure 3a) and average reliability score of online news consumed (Figure 3b) for each respondent with whom we have collected digital trace data. The pre-treatment value is plotted on the x-axis and the value between July 1–13 is plotted on the y-axis. In each figure, a simple linear regression line is plotted that predicts the value of each variable after the NewsGuard extension was disabled as a function of its pre-treatment value. In Figures 3a and 3b, the steeper slopes of the linear regression lines from the control groups show that improvement in news consumption quality due to the intervention is greatest among those who already consumed the most unreliable news on average.



The next hypothesis predicted that source reliability feedback would increase trust in the media and reliable sources (H2). Interestingly, we find that exposure to the treatment increases trust in media, but does not increase support in specific reliable news sources. The treatment effect estimates for each pre-registered attitudinal measure are presented in Figure 4. Figure 4 reveals a positive effect on trust in media ($\beta = 0.056$, SE $= 0.0279$, P < 0.05). Given that the *trust in media* score is an index from 0 to 3 with an average score of 1.63, this represents a 3.4% increase over the course of an approximately two-week intervention period.[12] Although we find an increase in trust in media generally, we do not find statistically reliable evidence that this treatment affected trust in selected specific sources deemed reliable by NewsGuard.

[12]More details on how this index was created can be found in the Supplementary Materials and Methods, Section A1.

Figure 4: This figure presents estimates of the effect of the intervention (with 95% confidence intervals) on our pre-registered attitudinal measures. The effect is reported in standard deviations of that measure (pre-treatment).



We find no support for our final hypothesis, which predicted that exposure to the treatment could help to alleviate pathologies such as affective polarization and political cynicism associated with consuming, believing, and sharing news from unreliable sources. RQ1 asked whether source reliability information affects belief in misinformation and true claims. To answer this question, all respondents were asked to judge the

veracity of five widely circulated statements about the Black Lives Matter (BLM) movement and five similarly well-circulated statements about Covid-19 using a four-point scale in Wave 2.Of the five statements about each topic, three were false and two were true. By taking the mean of the perceived veracity measure for the three false statements about each topic we create a measure for belief in misinformation in the BLM movement and Covid-19. By taking the mean of the perceived veracity measure for the two true statements about each topic we create a measure for belief in true information. Figure 4 shows that the intervention did in fact *reduce* belief in misinformation about the Black Lives Matter movement and misinformation about Covid-19, while at the same time not measurably affecting belief in true information. More specifically, we estimate an 8% decrease in perceived accuracy of BLM misinformation and a 13% decrease in perceived accuracy of Covid-19 misinformation.

RQ2 asked whether exposure to this intervention leads to downstream effects on outcomes such as trust in institutions, belief that fake news is a problem in general, and belief that fake news is a problem in the mainstream media. Figure 4 shows that the intervention does not affect these outcomes, with the exception of the belief that fake news is a problem in the mainstream media. Finally, RQ3 asked if any of the effects that we reported above (proportion of news consumed that is unreliable, reliability score of news consumed, trust in media, belief in misinformation, belief that fake news is a problem in the mainstream media) are moderated by specific characteristics. Although we find that the increase in media trust is concentrated among those that consume the highest proportion of unreliable news sources, we find no other consistent evidence of effect heterogeneity; the results from these models are presented in the Supplementary Materials and Methods in sections C4 and D4.

Discussion

Our analyses reveal that dynamic, in-feed source reliability information can lower the proportion of unreliable news consumed online by nearly 36%, reduce belief in misinformation without measurably reducing belief in contemporaneous accurate information, and increase trust in the media overall. Scientifically, our findings indicate that online news and information diets are not immutable and that contextual factors such as online choice architecture may have a larger role in shaping online information diets than previously understood (Lorenz-Spreen et al. 2020; Guess 2021). Moreover, we provide rare evidence on the behavioral impact of source quality information (Pennycook and Rand 2019a).

In light of these findings, it seems likely that policies and strategies designed to mitigate the consumption and belief in misinformation would benefit from incorporating similarly designed source reliability labels. To our knowledge, no other intervention has been shown to reduce actual consumption of unreliable news (Guess

et al. 2020b). Hypothetically, generalizing our estimates of treatment effects among compliers to the online population could lead to downstream consequences for publishers as well, as the demand for news and information from unreliable outlets could decrease. A sizable drop in traffic, and the resulting revenue loss, could remove some of the financial incentives for producing misinformation as well as deter future entrants into the market.

In addition to its mitigating effect on consumption of misinformation, our randomized encouragement to use NewsGuard appeared to change attitudes and beliefs about the media writ large, namely increasing trust in the media and reducing belief that "fake news" is a problem in the mainstream media. The combination of increased trust in media with lower consumption of unreliable news could even initiate a self-sustaining cycle: As online news consumers become more trusting of authoritative sources, they could become even less likely to consult unreliable news sources, which in turn could feed back into greater trust of established media. Such dynamics potentially explain why measurable effects persist after the browser extension no longer provided dynamic source feedback to our participants.

We show that news quality labels can reduce belief in misinformation without suffering from the negative spillovers or scalability challenges of other strategies. In particular, while we focus on a commonly known browser extension that offers well-validated source reliability ratings, expert judgments are not a requirement of this approach. On the contrary, prior research has shown that informative signals of source quality can be produced relatively cheaply by politically balanced samples of lay citizens (Pennycook and Rand 2019a). While our effect sizes are all relatively modest in magnitude, the period in which participants in the treatment group received visible source reliability labels was short: Participants were exposed to news source quality labels for only 2–3 weeks and responded to the Wave 2 survey on average 2 weeks after the treatment was assigned.Given evidence that our treatment effects grew over time, we consider our estimates a likely lower bound on the magnitude of possible impacts.

Our results contribute to a growing body of evidence suggesting that people prefer to consume higher-quality information, all else equal (Pennycook and Rand 2019b; Pennycook et al. 2021). Since sifting through the crowded information environment costs effort, well-designed cognitive shortcuts providing easy-to-understand contextual information about source reliability cause an overall improvement in news diet quality. Consistent with this account, our effects are more concentrated among those who consume the most information from unreliable sources. The cumulative nature of these effects, along with our survey-based results documenting more favorable attitudes and beliefs about the media, suggest a learning process in which exposure to higher-quality news and information can influence future consumption habits. Simple contextual information about news sources, like those evaluated here, could be incorporated as a browser feature or directly into the design of social media platforms and search engines. Our estimates of treatment

effects among compliers, which most directly shed light on this counterfactual policy regime, imply that news diets could be improved without sacrificing user autonomy or further increasing the public's skepticism toward reliable information sources.

Materials and Methods

NewsGuard Extension and Ratings

To produce credibility ratings, NewsGuard employs a team of trained journalists and editors to review and rate news and information websites based on nine journalistic criteria. The criteria assess basic practices of reliability and transparency. Based on a site's performance on these nine criteria, it is assigned a reliability rating from 0 and 100. Online domains with score of 60 or higher are considered reliable (green shield), while scores below 60 are considered unreliable.[13] A histogram of NewsGuard scores for the majority of online news domains can be found in Section B4 of the Supplementary Materials.[14] NewsGuard can be installed on all major web browsers (Safari, Microsoft Edge, Mozilla Firefox, Internet Explorer, and Google Chrome) as well as Android and iOS mobile phones. Normally, the NewsGuard extension costs $2.99 per month, but it is available for free (and bundled) with Microsoft Edge as well as to over 200 million potential users worldwide through assorted partnerships.[15]

Data Collection

We conducted a two-wave panel survey of respondents that included an encouragement to install NewsGuard. This two-wave online panel survey was fielded by the survey company YouGov in the summer of 2020 (Wave 1: May 28–June 9, $N = 3,862$; Wave 2: June 19–July 1, $N = 3,337$). Respondents were selected by YouGov's matching and weighting algorithm to approximate the demographic and political attributes of the U.S. population (32% college graduates, 45% male, median age 50 years old; 46% identify as Democrats and 36% as Republicans). We also oversampled members of the YouGov Pulse panel, who voluntarily provide behavioral data on their online information consumption ($N = 939$) (see Supplementary Materials and Methods, section B1 for demographic details). Pulse panelists confidentially share visit-level data on domains and URLs of web activity, including estimated duration and time stamps, on registered desktop/laptop and mobile devices. For thorough validation of Pulse data, see Guess et al. (2020a); Guess

[13] Over 41% of the more than 5,000 news domains rated received a red, suspect rating.

[14] Over the course of 2020, NewsGuard rated 2,144 additional news domains and partnered with the World Health Organization to report misinformation and flag 371 websites that spread misinformation about Covid-19 in the first months of the pandemic.

[15] Currently, NewsGuard is offered for free to 30 million BT internet and mobile customers, students through TurnItIn, and patrons at over 750 libraries (including the Chicago Public Library).

(2021).

Main Dependent Variable: Online Consumption of Unreliable News

The main outcome of interest is news consumed by our study participants from publishers of low quality news sources. Using digital trace data we measure this phenomenon with five different strategies for each respondent in three separate periods of the study: the period before they took the Wave 1 survey ; the period between their completion of the Wave 1 survey and June 30th ;[16] and the two-week period after June 30, 2020. On July 1st, NewsGuard transitioned from offering its extension for free to a cost of $2.99 a month. Given this switch we can assume that the vast majority of respondents were no longer using the NewsGuard extension after June 30th. During these three periods we calculate the average reliability score for websites visited, proportion of unreliable news sites visited, count of unreliable news sites visited, proportion of reliable news sites visited, and count of reliable news sites visited. The first measure, average reliability score, was derived by calculating the average NewsGuard reliability rating (0 to 100) of all news domains visited by the participant in that period. For the others measures, we labeled all news domains visited by that respondent in each period as unreliable (the domain has a reliability score from NewsGuard of below 60) or reliable (the domain has a reliability score from NewsGuard of 60 or above) and calculated the proportion and count of unreliable and reliable online news domains over the three periods of interest.

We are also interested in the effect of this intervention on the dependent variables specified in our other hypotheses and research questions, including the perceived accuracy of true and false news stories, trust in media, and other possible downstream effects. Details on how these variables are measured are available in the Supplementary Materials and Methods in section A1. We are also interested if effects on these variables are higher within groups most vulnerable to misinformation, such as those those that use social media more or those that have lower levels of digital literacy. Details on how these variables are measured can be found in the Supplementary Materials and Methods in section A2.

Analysis

Our pre-registered primary analyses are an Intent-to-Treat (ITT) model and two Complier Average Causal Effect (CACE) models using two different compliance measures, one that measures compliance solely at the beginning of the treatment period and another which measures compliance both at the beginning and end of the treatment period. We report both unadjusted (differences in means) and covariate-adjusted estimates of treatment effects for each dependent variable of interest.[17] For covariate-adjusted models, we

[16] The NewsGuard extension's free capabilities terminated on June 30, 2020, coinciding with the end of the treatment period.
[17] We use robust standard errors (HC2) in all analyses and report p-values from two-tailed t-tests.

selected covariates for inclusion using lasso run separately for each dependent variable.

The key explanatory variable of interest is exposure to the NewsGuard web extension. At the beginning of the survey, respondents in both the treatment and control group are asked if they would be willing to install an extension to their web browser, which acts to to minimize differences across the treatment and control group and in attrition. We then randomly assign respondents in Wave 1 to be encouraged to install the NewsGuard web extension. We do not find that those in the treatment and control groups were statistically different across income, race, partisanship, education, and gender (Demographic details are presented in the Supplementary Materials and Methods in section B1). Those in the treatment group were slightly younger (by 2 years) and had higher levels of digital literacy than the control group (by about one point on a 1 to 66 scale), but the magnitudes of these differences are small. We also found attrition rates between the treatment group and the control group to be similar. In fact the attrition rate in the control group was slightly higher than in the treatment group (14.1% in the control group compared to 13.2% in the treatment group). We do not find that those who attrited by not taking the Wave 2 survey in the control and treatment groups were statistically different in terms of income, race, partisanship, education, gender, age, or digital literacy (Details are presented in the Supplementary Materials and Methods in section B2).

The treatment was administered at the beginning of the Wave 1 survey. We define "compliance" as successfully installing and activating the NewsGuard extension (as a result of the encouragement), which we validate via a script linked at the beginning of the Wave 1 survey and during the last week of the treatment period.[18] This gives us two separate compliance measures that we can use for over 92% of our respondents.[19] Of those from which we have data from, 95% passed the first compliance check and 80% passed both the first and second compliance check. Notably, we find little difference in the characteristics of respondents who would successfully take the treatment if encouraged ("compliers")[20] and those who would not take the treatment if encouraged ("never takers"). We find no statistically significant evidence that respondents who comply differ in partisan leaning, education level, gender, race, or income-level. Compliers were more likely to be older, scored higher on our digital literacy scale, and consumed more unreliable news domains than never takers in the pre-treatment period, but the magnitudes of these differences are small.[21]

Measuring compliance is necessary to properly estimate the effects of this intervention. One concern

[18] Respondents click the verification link in both compliance checks and they are redirected to a separate page in which we can verify whether the NewsGuard extension has been installed and is active. We record their unique ID and the result of the compliance check, so we can match it to their survey responses. We did not ask early respondents of our Wave 2 survey to complete the compliance check during the survey. Rather, we waited until the last week of the treatment period and sent them an e-mail asking them to click on the verification link. Respondents who filled out the Wave 2 survey in the last week of the treatment period were asked to click on the verification link at the end of their Wave 2 survey.

[19] This compliance check failed about 4% of the time due to random browser or survey issues that do not appear biased in any identifiable direction. Given this, we only collected first and second compliance check data for 92% of our respondents

[20] In this analysis we define compliance using our second, stronger compliance check

[21] Details comparing these two groups can be found in the Supplementary Materials and Methods Section B3

with using an intent-to-treat model is that it will understate the true effect of an intervention when some respondents do not comply with the encouragement. We can offer the opportunity to install and activate the NewsGuard extension to a random subset of respondents, but we cannot force every respondent to do this. Our case is also unique in that some respondents in our control group (1%) are already compliant to the treatment and have the NewsGuard extension installed and activated. To account for varying levels of treatment within the treatment and control group we estimate the effect of the treatment on those who are actually treated. This is known as a Complier Average Causal Effect (CACE) model, which uses an instrumental variables approach. We instrument receipt of the NewsGuard treatment (compliance measure) with an indicator for treatment assignment (random assignment to the treatment group). In this paper we strictly present results from the covariate-adjusted[22] CACE model utilizing the stronger, second compliance check, which best measures the effect of this intervention.[23]

[22]For covariate-adjusted models (see above), we select covariates for inclusion using lasso with default options in `glmnet` (R) and seed set to 938. We run this procedure separately for each dependent variable, and use the same selected variables for both OLS (ITT) and 2SLS (CACE) models. Our list of pre-treatment covariates for possible inclusion can be found in the Supplementary Materials and Methods Section A3 and page 16. For moderation analyses, we use covariates selected for main effect then add the interaction (and base terms if necessary).

[23]We present the results from the covariate-adjusted and unadjusted intent-to-treat model and both CACE models for all analyses reported in the paper in the Supplementary Materials and Methods in Section C and D. The estimated effects of exposure to the treatment are consistently smaller in the ITT and first CACE model, but this is unsurprising as these models likely understate the effect of the treatment.

References

Anspach, N. M., Jennings, J. T., and Arceneaux, K. (2019). A little bit of knowledge: Facebook's news feed and self-perceptions of knowledge. *Research & Politics*, 6(1):2053168018816189.

Bakshy, E., Messing, S., and Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132.

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

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., et al. (2019). 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*, pages 1–23.

Coppock, A., Ekins, E., Kirby, D., et al. (2018). The long-lasting effects of newspaper op-eds on public opinion. *Quarterly Journal of Political Science*, 13(1):59–87.

Eady, G., Bonneau, R., Tucker, J. A., and Nagler, J. (2020). News Sharing on Social Media: Mapping the Ideology of News Media Content, Citizens, and Politicians.

Ecker, U. K., Lewandowsky, S., and Tang, D. T. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & cognition*, 38(8):1087–1100.

Fletcher, R., Cornia, A., and Nielsen, R. K. (2020). How polarized are online and offline news audiences? a comparative analysis of twelve countries. *The International Journal of Press/Politics*, 25(2):169–195.

Flynn, D., Nyhan, B., and Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, 38:127–150.

Gentzkow, M. and Shapiro, J. M. (2011). Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4):1799–1839.

Gerber, A. S., Gimpel, J. G., Green, D. P., and Shaw, D. R. (2011). How large and long-lasting are the persuasive effects of televised campaign ads? results from a randomized field experiment. *American Political Science Review*, pages 135–150.

Gigerenzer, G. and Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT press.

Gigerenzer, G. E., Hertwig, R. E., and Pachur, T. E. (2011). *Heuristics: The foundations of adaptive behavior*. Oxford University Press.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. (2019). Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378.

Guess, A., Nagler, J., and Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science advances*, 5(1):eaau4586.

Guess, A., Nyhan, B., and Reifler, J. (2020a). Exposure to untrustworthy websites in the 2016 us election. *Nature Human Behaviour*, 4(5):472–480.

Guess, A. M. (2021). (Almost) everything in moderation: New evidence on americans' online media diets. *American Journal of Political Science*.

Guess, A. M., Barberá, P., Munzert, S., and Yang, J. (2021). The consequences of online partisan media. *Proceedings of the National Academy of Sciences*.

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., and Sircar, N. (2020b). 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.

Iyengar, S. and Hahn, K. S. (2009). Red media, blue media: Evidence of ideological selectivity in media use. *Journal of communication*, 59(1):19–39.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., and Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C. R., and Hertwig, R. (2020). How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature Human Behaviour*, 0:1–8.

Lupia, A. (1994). Shortcuts versus encyclopedias: Information and voting behavior in california insurance reform elections. *American Political Science Review*, pages 63–76.

Messing, S. and Westwood, S. J. (2014). Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication research*, 41(8):1042–1063.

Munzert, S., Barberá, P., Guess, A., and Yang, J. (2021a). Do online voter guides empower citizens? evidence from a field experiment with digital trace data. *Public Opinion Quarterly*.

Munzert, S., Selb, P., Gohdes, A., Stoetzer, L. F., and Lowe, W. (2021b). Tracking and promoting the usage of a covid-19 contact tracing app. *Nature Human Behaviour*, pages 1–9.

Nyhan, B., Reifler, J., and Ubel, P. A. (2013). The hazards of correcting myths about health care reform. *Medical care*, pages 127–132.

Pennycook, G., Bear, A., Collins, E. T., and Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, 66(11):4944–4957.

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*.

Pennycook, G. and Rand, D. G. (2019a). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7):2521–2526.

Pennycook, G. and Rand, D. G. (2019b). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50.

Pennycook, G. and Rand, D. G. (2021). Examining false beliefs about voter fraud in the wake of the 2020 presidential election. *PsyArXiv Preprints*.

Sunstein, C. R. (2017). *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.

Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

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