Understanding and reducing the spread of misinformation online

Preprint · November 2019

DOI: 10.31234/osf.io/3n9u8

6 authors, including:

Gordon Pennycook
University of Regina
150 PUBLICATIONS **13,067** CITATIONS
SEE PROFILE

Mohsen Mosleh
Massachusetts Institute of Technology
43 PUBLICATIONS **730** CITATIONS
SEE PROFILE

Article

Shifting attention to accuracy can reduce misinformation online

https://doi.org/10.1038/s41586-021-03344-2

Received: 5 March 2020

Accepted: 8 February 2021

Published online: 17 March 2021

Check for updates

Gordon Pennycook[1,8,9]✉, Ziv Epstein[2,3,9], Mohsen Mosleh[3,4,9], Antonio A. Arechar[3,5], Dean Eckles[3,6] & David G. Rand[3,6,7]✉

In recent years, there has been a great deal of concern about the proliferation of false and misleading news on social media[1–4]. Academics and practitioners alike have asked why people share such misinformation, and sought solutions to reduce the sharing of misinformation[5–7]. Here, we attempt to address both of these questions. First, we find that the veracity of headlines has little effect on sharing intentions, despite having a large effect on judgments of accuracy. This dissociation suggests that sharing does not necessarily indicate belief. Nonetheless, most participants say it is important to share only accurate news. To shed light on this apparent contradiction, we carried out four survey experiments and a field experiment on Twitter; the results show that subtly shifting attention to accuracy increases the quality of news that people subsequently share. Together with additional computational analyses, these findings indicate that people often share misinformation because their attention is focused on factors other than accuracy—and therefore they fail to implement a strongly held preference for accurate sharing. Our results challenge the popular claim that people value partisanship over accuracy[8,9], and provide evidence for scalable attention-based interventions that social media platforms could easily implement to counter misinformation online.

The sharing of misinformation on social media—including, but not limited to, blatantly false political 'fake news' and misleading hyperpartisan content—has become a major focus of public debate and academic study in recent years[1,4]. Although misinformation is nothing new, the topic gained prominence in 2016 after the US Presidential Election and the UK's Brexit referendum, during which entirely fabricated stories (presented as legitimate news) received wide distribution via social media—a problem that has gained even more attention during the COVID-19 pandemic[2,7] and the Capitol Hill riot following the 2020 US Presidential Election[10].

Misinformation is problematic because it leads to inaccurate beliefs and can exacerbate partisan disagreement over even basic facts. Merely reading false news posts—including political posts that are extremely implausible and inconsistent with one's political ideology— makes them subsequently seem more true[11].

In addition to being concerning, the widespread sharing of misinformation on social media is also surprising, given the outlandishness of much of this content. Here we test three competing theories of why people share misinformation, based respectively on (i) confusion about what is (in)accurate, (ii) preferences for factors such as partisanship over accuracy, and (iii) inattention to accuracy.

Disconnect between sharing and accuracy

We begin with the confusion-based account, in which people share misinformation because they mistakenly believe that it is accurate (for example, owing to media or digital illiteracy[5,12–15] or politically motivated reasoning[8,9,16,17]). To gain initial insight into whether mistaken beliefs are sufficient to explain the sharing of misinformation, study 1 tests for a dissociation between what people deem to be accurate and what they would share on social media. We recruited $n = 1{,}015$ American individuals using Amazon Mechanical Turk (MTurk)[18], and presented them with the headline, lede sentence, and image for 36 actual news stories taken from social media. Half of the headlines were entirely false and half were true; half of the headlines were chosen (via pretest[19,20]) to be favourable to Democrats and the other half to be favourable to Republicans. Participants were randomly assigned to then either judge the veracity of each headline (accuracy condition) or indicate whether they would consider sharing each headline online (sharing condition) (for details, see Methods). Unless noted otherwise, all P values are generated by linear regression with robust standard errors clustered on participant and headline.

In the accuracy condition (Fig. 1a), true headlines were rated as accurate significantly more often than false headlines (55.9 percentage point difference, $F_{(1, 36{,}172)} = 375.05$, $P < 0.0001$). Although politically concordant headlines were also rated as accurate significantly more often than politically discordant headlines (10.1 percentage point difference, $F_{(1, 36{,}172)} = 26.45$, $P < 0.0001$), this difference based on partisan alignment was significantly smaller than the 55.9 percentage point veracity-driven difference between true and false headlines ($F_{(1, 36{,}172)} = 137.26$, $P < 0.0001$). Turning to the sharing condition (Fig. 1b), we see the opposite pattern. Whether the headline was politically

[1]Hill/Levene Schools of Business, University of Regina, Regina, Saskatchewan, Canada. [2]Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA. [4]Department of SITE (Science, Innovation, Technology, and Entrepreneurship), University of Exeter Business School, Exeter, UK. [5]Center for Research and Teaching in Economics (CIDE), Aguascalientes, Mexico. [6]Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA, USA. [7]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [8]Department of Psychology, University of Regina, Regina, Saskatchewan, Canada. [9]These authors contributed equally: Gordon Pennycook, Ziv Epstein, Mohsen Mosleh. ✉e-mail: gordon.pennycook@uregina.ca; drand@mit.edu

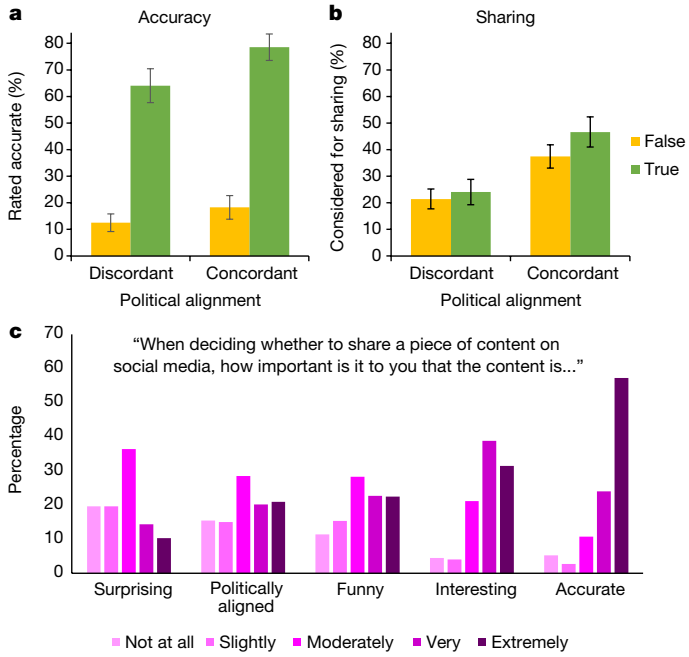


Fig. 1 | Sharing intentions are much less discerning than accuracy judgements—despite an overall desire to share only accurate content. In study 1, $n = 1{,}002$ American individuals from Amazon MTurk were presented with a set of 36 headlines and asked to indicate whether they thought the headlines were accurate or whether they would consider sharing them on social media. **a**, The fraction of headlines rated as accurate in the 'accuracy' condition, by the veracity of the headline and political alignment between the headline and the participant. Participants were significantly more likely to rate true headlines as accurate compared to false headlines (55.9 percentage point difference, $F_{(1, 36,172)} = 375.05$, $P < 0.0001$), whereas the partisan alignment of the headlines had a significantly smaller effect (10.1 percentage point difference, $F_{(1, 36,172)} = 26.45$, $P < 0.0001$; interaction, $F_{(1, 36,172)} = 137.26$, $P < 0.0001$). **b**, The fraction of headlines that participants said they would consider sharing in the 'sharing' condition, by the veracity of the headline and political alignment between the headline and the participant. In contrast to the accuracy condition, the effect of headline veracity was significantly smaller in the sharing condition, $F_{(1, 36,172)} = 260.68$, $P < 0.0001$, whereas the effect of political concordance was significantly larger, $F_{(1, 36,172)} = 17.24$, $P < 0.0001$. Error bars indicate 95% confidence intervals based on robust standard errors clustered on participant and headline, and all P values were generated using linear regression with robust standard errors clustered on participant and headline. **c**, Participants nonetheless overwhelmingly said that they thought that accuracy was more important on average than partisanship (and all other content dimensions that we asked about) when making decisions about what to share on social media (data shown from study 2; for study 1 see Extended Data Fig. 1).

concordant or discordant had a significantly larger effect on sharing intentions (19.3 percentage points) than whether the headline was true or false (5.9 percentage points; $F_{(1, 36,172)} = 19.73$, $P < 0.0001$). Accordingly, the effect of headline veracity was significantly larger in the accuracy condition than in the sharing condition ($F_{(1, 36,172)} = 260.68$, $P < 0.0001$), whereas the effect of concordance was significantly larger in the sharing condition than in the accuracy condition ($F_{(1, 36,172)} = 17.24$, $P < 0.0001$; for the full regression table and robustness checks, see Supplementary Information section 2). Notably, the pattern of sharing intentions that we observe here matches the pattern of actual sharing observed in a large-scale analysis of Twitter users, in which partisan alignment was found to be a much stronger predictor of sharing than veracity[21].

To illustrate the disconnect between accuracy judgments and sharing intentions, consider, for example, the following headline: 'Over 500 'Migrant Caravaners' Arrested With Suicide Vests'. This was rated as accurate by 15.7% of Republicans in our study, but 51.1% of Republicans said they would consider sharing it. Thus, the results from study 1 suggest that the confusion-based account cannot fully explain the sharing of misinformation: our participants were more than twice as likely to consider sharing false but politically concordant headlines (37.4%) as they were to rate such headlines as accurate (18.2%; $F_{(1, 36,172)} = 19.73$, $P < 0.0001$).

One possible explanation for this dissociation between accuracy judgments and sharing intentions is offered by the preference-based account of misinformation sharing. By this account, people care about accuracy much less than other factors (such as partisanship), and therefore knowingly share misinformation. The fact that participants in study 1 were willing to share ideologically consistent but false headlines could thus be reasonably construed as revealing their preference for weighting non-accuracy dimensions (such as ideology) over accuracy. Yet when asked at the end of the study whether it is important to share only content that is accurate on social media, the modal response was "extremely important" (Extended Data Fig. 1). A similar pattern was observed in a more nationally representative sample of $n = 401$ American individuals from Lucid[22] in study 2, who rated accuracy as substantially more important for social media sharing than any of the other dimensions that we asked about (paired t-tests, $P < 0.001$ for all comparisons) (Fig. 1c; for design details, see Methods).

Why, then, were the participants in study 1—along with millions of other American people in recent years—willing to share misinformation? In answer, we advance the inattention-based account, in which (i) people do care more about accuracy than other content dimensions, but accuracy nonetheless often has little effect on sharing, because (ii) the social media context focuses their attention on other factors such as the desire to attract and please followers/friends or to signal one's group membership[23–25]. In the language of utility theory, we argue that an 'attentional spotlight' is shone upon certain terms in the decider's utility function, such that only those terms are weighed when making a decision (for a mathematical formalization of this limited-attention utility model, see Supplementary Information section 3).

Priming accuracy improves sharing

We differentiate between these theories by subtly inducing people to think about accuracy, which the preference-based account predicts should have no effect whereas the inattention-based account predicts should increase the accuracy of content that is shared (see Supplementary Information section 3.2). We first test these competing predictions by performing a series of survey experiments with similar designs. In the control condition of each experiment, participants were shown 24 news headlines (balanced on veracity and partisanship, as in study 1) and asked how likely they would be to share each headline on Facebook. In the treatment condition, participants were asked to rate the accuracy of a single non-partisan news headline at the outset of the study (ostensibly as part of a pretest for stimuli for another study). They then went on to complete the same sharing intentions task as in the control condition, but with the concept of accuracy more likely to be salient in their minds. For details of the experimental design, see Methods.

In two experiments using American individuals recruited from MTurk (study 3, $n = 727$; study 4, $n = 780$), we find that the treatment condition significantly increased sharing discernment (interaction between headline veracity and treatment: study 3, $b = 0.053$, 95% confidence interval [0.032, 0.074], $F_{(1, 17,413)} = 24.21$, $P < 0.0001$; study 4, $b = 0.065$, 95% confidence interval [0.036, 0.094], $F_{(1, 18,673)} = 19.53$, $P < 0.0001$) (Fig. 2a, b). Specifically, participants in the treatment group were significantly less likely to consider sharing false headlines compared to those in the control group (study 3, $b = -0.055$, 95% confidence interval [−0.083, −0.026], $F_{(1, 17,413)} = 14.08$, $P = 0.0002$; study 4, $b = -0.058$, 95% confidence interval [−0.091, −0.025], $F_{(1, 18,673)} = 11.99$, $P = 0.0005$), but equally likely to consider sharing true headlines (study 3, $b = -0.002$, 95% confidence interval [−0.031, 0.028], $F_{(1, 17,413)} = 0.01$, $P = 0.92$; study 4, $b = 0.007$, 95% confidence interval [−0.020, 0.033],

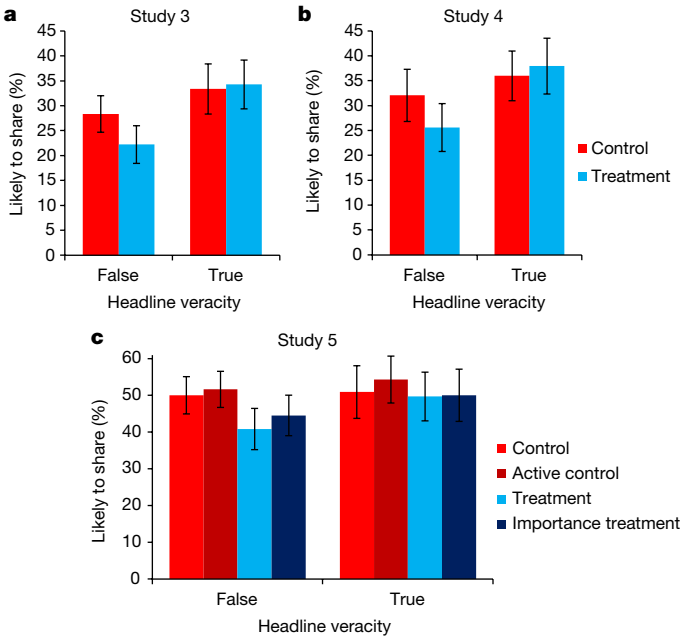


Fig. 2 | Inducing survey respondents to think about accuracy increases the veracity of headlines they are willing to share. a–c, Participants in study 3 (**a**; $n = 727$ American individuals from MTurk), study 4 (**b**; $n = 780$ American individuals from MTurk) and study 5 (**c**; $n = 1{,}268$ American individuals from Lucid, quota-matched to the national distribution on age, gender, ethnicity and geographical region) indicated how likely they would be to consider sharing a series of actual headlines from social media. Participants in the 'treatment' condition rated the accuracy of a single non-political headline at the outset of the study, thus increasing the likelihood that they would think about accuracy when indicating sharing intentions relative to the 'control' condition. In study 5, we added an 'active control' (in which participants rated the humorousness of a single headline at the outset of the study) and an 'importance treatment' (in which participants were asked at the study outset how important they thought it was to share only accurate content). For interpretability, shown here is the fraction of 'likely' responses (responses above the midpoint of the six-point Likert scale) by condition and headline veracity; the full distributions of responses are shown in Extended Data Figs. 2, 3. As per our preregistered analysis plans, these analyses focus only on participants who indicated that they sometimes consider sharing political content on social media; for analysis including all participants, see Supplementary Information section 2. Error bars indicate 95% confidence intervals based on robust standard errors clustered on participant and headline.

$F_{(1,\ 18,673)} = 0.23$, $P = 0.63$). As a result, sharing discernment (the difference in sharing intentions for true versus false headlines) was 2.0 times larger in the treatment relative to the control group in study 3, and 2.4 times larger in study 4. Furthermore, there was no evidence of a backfire effect, as the treatment effect was actually significantly larger for politically concordant headlines than for politically discordant headlines ($b = 0.022$, 95% confidence interval [0.012, 0.033], $F_{(1,\ 36,078)} = 18.09$, $P < 0.0001$), and significantly increased discernment for both Democrats ($b = 0.069$, 95% confidence interval [0.048, 0.091], $F_{(1,\ 24,636)} = 40.38$, $P < 0.0001$) and Republicans ($b = 0.035$, 95% confidence interval [0.007, 0.063], $F_{(1,\ 11,394)} = 5.93$, $P = 0.015$). See Supplementary Information section 2 for the full regression table.

Notably, there was no significant difference between conditions in responses to a post-experimental question about the importance of sharing only accurate content (*t*-test: $t_{(1498)} = 0.42$, $P = 0.68$, 95% confidence interval [−0.075, 0.115] points on a 1–5 scale; Bayesian independent samples *t*-test with Cauchy prior distribution with interquartile range of 0.707: $BF_{10} = 0.063$, providing strong evidence for the null), or regarding participants' perceptions of the importance that their friends place on sharing only accurate content (*t*-test: $t_{(768)} = -0.57$, $P = 0.57$, 95% confidence interval [−0.205, 0.113] points on a 1–5 scale; Bayesian independent samples *t*-test with Cauchy prior distribution with interquartile range of 0.707: $BF_{10} = 0.095$, providing strong evidence for the null).

Our next survey experiment (study 5, $n = 1{,}268$) tested whether the previous results generalize to a more representative sample by recruiting participants from Lucid[22] that were quota-sampled to match the distribution of American residents on age, gender, ethnicity and geographical region. Study 5 also included an active control condition in which participants were asked to rate the humorousness (rather than accuracy) of a single non-partisan news headline at the outset of the study, and an importance treatment condition that tested another approach for making accuracy salient by having participants begin the study by indicating the importance they place on sharing only accurate content (instead of rating the accuracy of a neutral headline). The results (Fig. 2c) successfully replicated studies 3 and 4. As expected, there were no significant differences in sharing intentions between the control and the active control conditions (interaction between veracity and condition, $b = 0.015$, 95% confidence interval [−0.043, 0.059], $F_{(1,\ 6,772)} = 0.04$, $P = 0.84$); and both treatments significantly increased sharing discernment relative to the controls (interaction between veracity and condition: treatment, $b = 0.054$, 95% confidence interval [0.023, 0.085], $F = 11.98$, $P = 0.0005$; importance treatment, $b = 0.038$, 95% confidence interval [0.014, 0.061], $F = 9.76$, $P = 0.0018$). See Supplementary Information section 2 for the full regression table.

Attending to accuracy as the mechanism

Next, we provide evidence that shifting attention to accuracy is the mechanism behind this effect by showing that the treatment condition leads to the largest reduction in the sharing of headlines that participants are likely to deem to be the most inaccurate (and vice versa for the most plainly accurate headlines). A headline-level analysis finds a positive correlation between the effect of the treatment on sharing and the headline's perceived accuracy (as measured in pre-tests, see Supplementary Information section 1) (study 3, $r_{(22)} = 0.71$, $P = 0.0001$; study 4, $r_{(22)} = 0.67$, $P = 0.0003$; study 5, $r_{(18)} = 0.61$, $P = 0.005$) (Fig. 3a–c). That is, the most obviously inaccurate headlines are the ones that the accuracy salience treatment most effectively discourages people from sharing.

Furthermore, fitting our formal limited-attention utility model to the experimental data provides quantitative evidence against the preference-based account (participants value accuracy as much as or more than partisanship) and for the inattention-based account (participants often do not consider accuracy) (Extended Data Table 1, Supplementary Information sections 3.5, 3.6).

In study 6, we carried out a final survey experiment ($n = 710$ American individuals from MTurk) that quantifies the relative contribution of the confusion-based, preference-based and inattention-based accounts to the willingness to share false headlines on social media. To do so, we compare the control condition to a 'full attention' treatment, in which participants are asked to assess the accuracy of each headline immediately before deciding whether they would share it (for details, see Methods). As illustrated in Fig. 3d, the results show that, of the sharing intentions for false headlines, the inattention-based account explains 51.2% (95% confidence interval [38.4%, 62.0%]) of sharing, the confusion-based account explains 33.1% (95% confidence interval [25.1%, 42.4%]) of sharing, and the preference-based account explains 15.8% (95% confidence interval [11.1%, 21.5%]) of sharing. Thus, inattention does not merely operate on the margin, but instead has a central role in the sharing of misinformation in our experimental paradigm. Furthermore, the preference-based account's low level of explanatory power relative to the inattention-based account in study 6 is consistent with the model fitting results in Extended Data Table 1 and Supplementary Information section 3.6 described above—thus providing

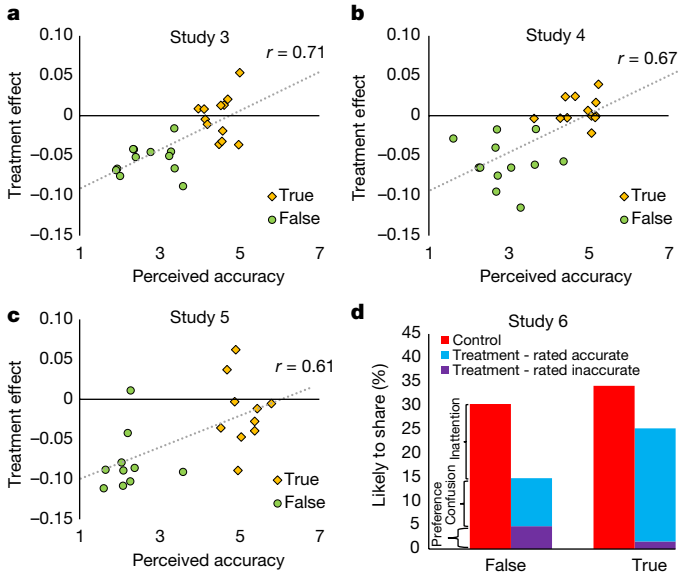


Fig. 3 | Inattention has an important role in the sharing of misinformation. **a**–**c**, There is a significant positive correlation across headlines between perceived accuracy (on the basis of pre-tests) and the effect of the treatment on sharing in study 3 (**a**, $r_{(22)} = 0.71$, $P = 0.0001$), study 4 (**b**, $r_{(22)} = 0.67$, $P = 0.0003$), and study 5 (**c**, $r_{(18)} = 0.61$, $P = 0.005$). The accuracy reminder caused a larger decrease in sharing intentions for items that were deemed to be more unlikely. This observation supports our argument that the treatment intervention operated via focusing attention on accuracy, and that many people do not want to share content that they think is inaccurate. As shown in Extended Data Fig. 4, in study 5 a similar pattern was found for the 'important treatment', and no such effect existed for the 'active control'. **d**, In study 6, participants rated the accuracy of each headline (a full-attention treatment) before making a judgment about sharing. This allows us to distinguish between false items that: (i) participants share despite believing to be inaccurate (that is, a preference-based rejection of truth); (ii) participants share and also believe to be accurate (that is, confusion-based); and (iii) participants no longer shared once they considered accuracy (that is, inattention-based). Results indicate that, among the false headlines that are shared in the control group, most are shared owing to inattention (51.2%), fewer are shared because of confusion (33.1%), and a small minority are shared because of a preference to share false content (15.8%). Bootstrapping simulations (10,000 repetitions) showed that inattention explains significantly more sharing than purposeful sharing ($b = 0.354$ [0.178, 0.502], $P = 0.0004$); that confusion explains significantly more sharing than purposeful sharing ($b = 0.173$ [0.098, 0.256], $P < 0.0001$); and that although inattention explained directionally more misinformation sharing than confusion, this difference was not statistically significant ($b = 0.181$ [−0.036, 0.365], $P = 0.098$).

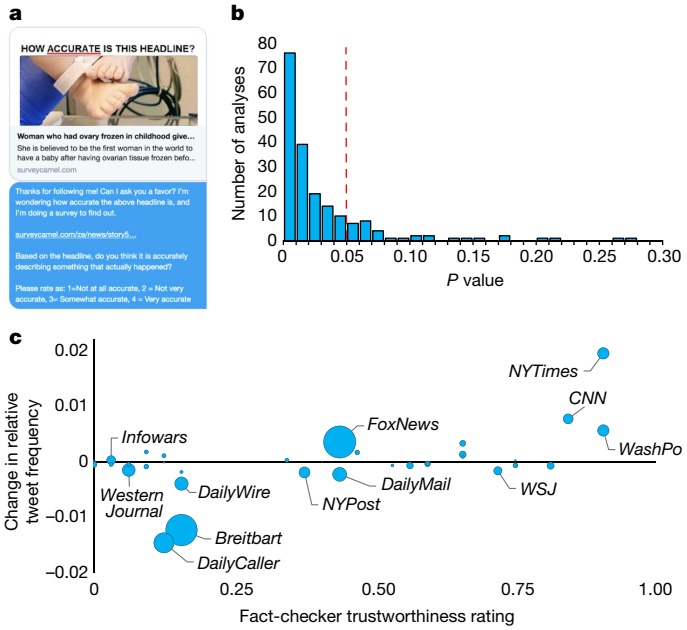


Fig. 4 | Sending Twitter users a message asking for their opinion about the accuracy of a single non-political headline increases the quality of the news that they subsequently share. In Study 7, we conducted an experiment on the Twitter platform involving $n = 5,379$ users who had recently shared links to websites that regularly produce misleading and hyperpartisan content. We randomized the date on which users were sent an unsolicited message asking them to rate the accuracy of a single non-political headline. We then compared the quality of the news sites shared in the 24 h after receiving the message to the sites shared by participants who had not yet received the message. **a**, The private message sent to the users is shown here. We did not expect most users to respond to the message, or even read it in its entirety. Thus, we designed it such that reading only the top line should be sufficient to shift attention to the concept of accuracy. **b**, To test the robustness of our results, we conducted 192 analyses that differed in their dependent variable, inclusion criteria and model specifications. Shown here is the distribution of *P* values resulting from each of these analyses. More than 80% of approaches yield $P < 0.05$. **c**, A domain-level analysis provides a more detailed picture of the effect of the intervention. The *x* axis indicates the trust score given to each outlet by professional fact-checkers[27]. The *y* axis indicates the fraction of rated links to each outlet in the 24 h after the intervention minus the fraction of links to each outlet among not-yet-treated users. The size of each dot is proportional to the number of pre-treatment posts with links to that outlet. Domains with more than 500 pre-treatment posts are labelled. *NYPost*, *New York Post*; *NYTimes*, *New York Times*; *WashPo*, *Washington Post*; *WSJ*, *Wall Street Journal*.

convergent evidence against the preference-based account being a central driver of misinformation sharing.

Deploying the intervention on Twitter

Finally, to test whether our findings generalize to natural social media use settings (rather than laboratory experiments), actual (rather than hypothetical) sharing decisions, and misinformation more broadly (rather than just blatant 'fake news'), in study 7 we conducted a digital field experiment on social media[26]. To do so, we selected $n = 5,379$ Twitter users who had previously shared links to two particularly well-known right-leaning sites that professional fact-checkers have rated as highly untrustworthy[27]: www.Breitbart.com and www.Infowars.com. We then sent these users private messages asking them to rate the accuracy of a single non-political headline (Fig. 4a). We used a stepped-wedge design to observe the causal effect of the message on the quality of the news content (on the basis of domain-level ratings of professional fact-checkers[27]) that the users shared in the 24 hours after receiving our intervention message. For details of the experimental design, see Methods.

Examining baseline (pre-treatment) sharing behaviour shows that we were successful in identifying users with relatively low-quality news-sharing habits. The average quality score of news sources from pre-treatment posts was 0.34. (For comparison, the fact-checker-based quality score was 0.02 for *Infowars*; 0.16 for *Breitbart*; 0.39 for *Fox News*, and 0.93 for the *New York Times*.) Moreover, 46.6% of shared news sites were sites that publish false or misleading content (0.9% fake news sites, 45.7% hyperpartisan sites).

Consistent with our survey experiments, we find that the single accuracy message made users more discerning in their subsequent sharing decisions (using Fisherian randomization inference[28] to calculate exact *P* values, P_{FRI}, based on the distribution of the *t* statistic under the null hypothesis). Relative to baseline, the accuracy message increased the average quality of the news sources shared ($b = 0.007$, $t_{(5375)} = 2.91$, 95% null acceptance region of *t* [−0.44, 2.59], $P_{FRI} = 0.009$) and the total quality of shared sources summed over all posts ($b = 0.014$,

$t_{(5375)} = 3.12$, 95% null acceptance region of t [−0.08, 2.90], $P_{FRI} = 0.011$). This translates into increases of 4.8% and 9.0%, respectively, when estimating the treatment effect for user-days on which tweets would occur in treatment (that is, excluding user-days in the 'never-taker' principal stratum[29,30], because the treatment cannot have an effect when no tweets would occur in either treatment or control); including user-days with no tweets yields an increase of 2.1% and 4.0% in average and total quality, respectively. Furthermore, the level of sharing discernment (that is, the difference in number of mainstream versus fake or hyperpartisan links shared per user-day; interaction between post-treatment dummy and link type) was 2.8 times higher after receiving the accuracy message ($b = 0.059$, $t_{(5371)} = 3.27$, 95% null acceptance region of t [−0.31, 2.67], $P_{FRI} = 0.003$).

To provide further support for the inattention-based account, we contrast lower-engagement sharing (in which the user simply re-shares content posted by another user: that is, retweets without comment) with higher-engagement sharing (in which the poster invests some time and effort to craft their own post or add a comment when sharing another post). Lower-engagement sharing, which accounts for 72.4% of our dataset, presumably involves less attention than higher-engagement sharing—therefore the inattention-based account of misinformation sharing predicts that our manipulation should primarily affect lower-engagement sharing. Consistent with this prediction, we observe a significant positive interaction ($b = 0.008$, $t_{(5371)} = 2.78$, 95% null acceptance region of t [−0.80, 2.24], $P_{FRI} = 0.004$), such that the treatment increases average quality of lower-engagement sharing but not higher-engagement sharing. Furthermore, we found no significant treatment effect on the number of posts without links to any of the news sites used in our main analyses ($b = 0.266$, $t_{(5375)} = 0.50$, 95% null acceptance region of t [−1.11, 1.64], $P_{FRI} = 0.505$).

Notably, the significant effects that we observed are not unique to one particular set of analytic choices. Figure 4b shows the distribution of P values observed in 192 different analyses assessing the overall treatment effect on average quality, summed quality, or discernment under a variety of analytic choices. Of these analyses, 82.3% indicate a significant positive treatment effect (and none of 32 analyses of posts without links to a news site—in which we would not expect a treatment effect—find a significant difference). For details, see Extended Data Table 4 and Supplementary Information section 5.

Finally, we examine the data at the level of the domain (Fig. 4c). We see that the treatment effect is driven by increasing the fraction of rated-site posts with links to mainstream new sites with strong editorial standards such as the *New York Times*, and decreasing the fraction of rated-site posts that linked to relatively untrustworthy hyperpartisan sites such as *Breitbart*. Indeed, a domain-level pairwise correlation between fact-checker rating and change in sharing due to the intervention shows a very strong positive relationship (domains weighted by number of pre-treatment posts; $r_{(44)} = 0.74$, $P < 0.0001$), replicating the pattern observed in the survey experiments (Fig. 3a–c). In summary, our accuracy message successfully induced Twitter users who regularly shared misinformation to increase the average quality of the news that they shared.

In Supplementary Information section 6, we use computational modelling to connect our empirical observations about individual-level sharing decisions in studies 3–7 to the network-level dynamics of misinformation spread. Across a variety of network structures, we observe that network dynamics can substantially amplify the magnitude of treatment effects on sharing (Extended Data Fig. 6). Improving the quality of the content shared by one user improves the content that their followers see, and therefore improves the content that their followers share. This in turn improves what the followers' followers see and share, and so on. Thus, the cumulative effects of such an intervention on how misinformation spreads across networks may be substantially larger than what is observed when only examining the treated individuals—particularly given that, in study 7, we find that the treatment is as effective, if not more so, for users with larger numbers of followers (see Supplementary Information section 5); and that our treatment effect size estimates in study 7 are conservative because we do not know when (or if) users actually saw our intervention message.

Conclusion

Together, these studies suggest that when deciding what to share on social media, people are often distracted from considering the accuracy of the content. Therefore, shifting attention to the concept of accuracy can cause people to improve the quality of the news that they share. Furthermore, we found a dissociation between accuracy judgments and sharing intentions that suggests that people may share news that they do not necessarily have a firm belief in. As a consequence, people's beliefs may not be as partisan as their social media feeds seem to indicate. Future work is needed to more precisely identify people's state of belief when not reflecting on accuracy. Is it that people hold no particular belief one way or the other, or that they tend to assume content is true by default[31]?

A substantial limitation of our studies is that they are focused on the sharing of political news among American individuals. In a recent set of follow-up survey experiments, our findings of a disconnect between accuracy and sharing judgments in study 1 and our treatment increasing sharing discernment in studies 3, 4 and 5 were successfully replicated using headlines about COVID-19 with quota-matched American samples[7]. Future work should examine applications to other content domains, including organized misinformation campaigns from political elites (such as about climate change[32] or fraud in the 2020 US Presidential Election[10]), and explore cross-cultural generalizability. Extending the Twitter field experiment design used in study 7 is also a promising direction for future work, including using a more continuous shock-based model of how (and when) the treatment affects individuals rather than the conservative intent-to-treat approach used here, examining more than 24 hours after the intervention, generalizing beyond users who follow-back experimenter accounts, testing an active control, and using article-level quality rather than domain-level quality scores.

Our results suggest that the current design of social media platforms—in which users scroll quickly through a mixture of serious news and emotionally engaging content, and receive instantaneous quantified social feedback on their sharing—may discourage people from reflecting on accuracy. But this need not be the case. Our treatment translates easily into interventions that social media platforms could use to increase users' focus on accuracy. For example, platforms could periodically ask users to rate the accuracy of randomly selected headlines, thus reminding them about accuracy in a subtle way that should avoid reactance[33] (and simultaneously generating useful crowd ratings that can help to identify misinformation[27,34]). Such an approach could potentially increase the quality of news circulating online without relying on a centralized institution to certify truth and censor falsehood.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03344-2.

1. Lazer, D. et al. The science of fake news. *Science* **9**, 1094–1096 (2018).
2. Lederer, E. UN chief says misinformation about COVID-19 is new enemy. *ABC News* https://abcnews.go.com/US/wireStory/chief-misinformation-covid-19-enemy-69850124 (accessed 4 April 2020).
3. Pasquetto, I. et al. Tackling misinformation: what researchers could do with social media data. *HKS Misinformation Review* **1**, 8 (2020).
4. Pennycook, G. & Rand, D. G. The psychology of fake news. *Trends. Cogn. Sci.* (in the press).

5. Guess, A. M. et al. A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proc. Natl Acad. Sci. USA* **117**, 15536–15545 (2020).
6. Kozyreva, A., Lewandowsky, S. & Hertwig, R. Citizens versus the internet: confronting digital challenges with cognitive tools. *Psychol. Sci. Public Interest* **21**, 103–156 (2020).
7. Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G. & Rand, D. G. Fighting COVID-19 misinformation on social media: experimental evidence for a scalable accuracy nudge intervention. *Psychol Sci.* **31**, 770–780 (2020).
8. Van Bavel, J. J. & Pereira, A. The partisan brain: an identity-based model of political belief. *Trends Cogn. Sci.* **22**, 213–224 (2018).
9. Kahan, D. M. *Misconceptions, Misinformation, and the Logic of Identity-Protective Cognition*. Cultural Cognition Project Working Paper Series No. 164, Yale Law School, Public Law Research Paper No. 605, Yale Law & Economics Research Paper No. 575. https://doi.org/10.2139/ssrn.2973067 (2017).
10. Pennycook, G. & Rand, D. G. Research note: Examining false beliefs about voter fraud in the wake of the 2020 Presidential Election. *HKS Misinformation Rev*. **2**, 1 (2021).
11. Pennycook, G., Cannon, T. D. & Rand, D. G. Prior exposure increases perceived accuracy of fake news. *J. Exp. Psychol. Gen*. **147**, 1865–1880 (2018).
12. McGrew, S., Ortega, T., Breakstone, J. & Wineburg, S. The challenge that's bigger than fake news: civic reasoning in a social media environment. *Am. Educ*. **41**, 4–9 (2017).
13. Lee, N. M. Fake news, phishing, and fraud: a call for research on digital media literacy education beyond the classroom. *Commun. Educ*. **67**, 460–466 (2018).
14. McDougall, J., Brites, M. J., Couto, M. J. & Lucas, C. Digital literacy, fake news and education. *Cult. Educ*. **31**, 203–212 (2019).
15. Jones-Jang, S. M., Mortensen, T. & Liu, J. Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. *Am. Behav. Sci*. **65**, 371–388 (2019).
16. Redlawsk, D. Hot cognition or cool consideration? Testing the effects of motivated reasoning on political decision making. *J. Polit*. **64**, 1021–1044 (2002).
17. Strickland, A. A., Taber, C. S. & Lodge, M. Motivated reasoning and public opinion. *J. Health Polit. Policy Law* **36**, 935–944 (2011).
18. Horton, J., Rand, D. & Zeckhauser, R. The online laboratory: conducting experiments in a real labor market. *Exp. Econ*. **14**, 399–425 (2011).
19. Pennycook, G. & Rand, D. G. Lazy, not biased: susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* **188**, 39–50 (2019).
20. Pennycook, G., Bear, A., Collins, E. & Rand, D. G. The implied truth effect: attaching warnings to a subset of fake news stories increases perceived accuracy of stories without warnings. *Manage. Sci*. **66**, 11 (2020).
21. Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. & Lazer, D. Fake news on twitter during the 2016 U.S. Presidential election. *Science* **363**, 374–378 (2019).
22. Coppock, A. & Mcclellan, O. A. Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Res. Polit*. https://doi.org/10.1177/2053168018822174 (2019).
23. Marwick, A. E. & Boyd, D. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media Soc*. **13**, 114–133 (2011).
24. Donath, J. & Boyd, D. Public displays of connection. *BT Technol. J*. **22**, 71–82 (2004).
25. Mosleh, M., Martel, C., Eckles, D. & Rand, D. G. Shared partisanship dramatically increases social tie formation in a Twitter field experiment. *Proc. Natl Acad. Sci. USA* **118**, e2022761118 (2021).
26. Munger, K. Tweetment effects on the tweeted: experimentally reducing racist harassment. *Polit. Behav*. **39**, 629–649 (2017).
27. Pennycook, G. & Rand, D. G. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc. Natl Acad. Sci. USA* **116**, 2521–2526 (2019).
28. Fisher, R. A. *The Design of Experiments* (Oliver and Boyd, 1937).
29. Angrist, J. D., Imbens, G. W. & Rubin, D. B. Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc*. **91**, 444 (1996).
30. Frangakis, C. E. & Rubin, D. B. Principal stratification in causal inference. *Biometrics* **58**, 21–29 (2002).
31. Gilbert, D. T. How mental systems believe. *Am. Psychol*. **46**, 107–119 (1991).
32. Dunlap, R. E. & McCright, A. M. *Organized Climate Change Denial* (eds Schlosberg, D. et al.) 144–160 (Oxford Univ. Press, 2011).
33. Mosleh, M., Martel, C., Eckles, D. & Rand, D. G. Perverse consequences of debunking in a twitter field experiment: being corrected for posting false news increases subsequent sharing of low quality, partisan, and toxic content. In *Proc. 2021 CHI Conference on Human Factors in Computing Systems* (2021).
34. Allen, J., Arechar, A. A., Pennycook, G. & Rand, D. G. Scaling up fact-checking using the wisdom of crowds. Preprint at https://doi.org/10.31234/osf.io/9qdza (2020).

Methods

Preregistrations for all studies are available at https://osf.io/p6u8k/. In all survey experiments, we do not exclude participants for inattentiveness or straightlining to avoid selection effects that can undermine causal inference. The researchers were not blinded to the hypotheses when carrying out the analyses. All experiments were randomized except for study 2, which was not randomized. No statistical methods were used to predetermine sample size.

Study 1

In study 1, participants were presented with a pretested set of false and true headlines (in 'Facebook format') and were asked to indicate either whether they thought they were accurate or not, or whether they would consider sharing them on social media or not. Our prediction was that the difference in 'yes' responses between false and true news (that is, discernment) will be greater when individuals are asked about accuracy than when they are asked about sharing, whereas the difference between politically discordant and concordant news will be greater when they are asked about sharing than when they are asked about accuracy.

Participants. We preregistered a target sample of 1,000 complete responses, using participants recruited from Amazon's MTurk but noted that we would retain individuals who completed the study above the 1,000-participant quota. In total, 1,825 participants began the survey. However, an initial (pre-treatment) screener only allowed American participants who indicated having a Facebook or Twitter account (when shown a list of different social media platforms) and indicated that they would consider sharing political content (when shown a list of different content types) to continue and complete the survey. The purpose of these screening criteria was to focus our investigation on the relevant subpopulation—those who share political news. The accuracy judgments of people who never share political news on social media are not relevant here, given our interest in the sharing of political misinformation. Of the participants who entered the survey, 153 indicated that they had neither a Facebook nor a Twitter account, and 651 indicated that they did have either a Facebook or Twitter account but would not consider sharing political content. A further 16 participants passed the screener but did not finish the survey and thus were removed from the dataset. The full sample (mean age = 36.7) included 475 males, 516 females, and 14 participants who selected another gender option. This study was run on 13–14 August 2019.

Materials. We presented participants with 18 false and 18 true news headlines in a random order for each participant. The false news headlines were originally selected from a third-party fact-checking website, www.Snopes.com, and were therefore verified as being fabricated and untrue. The true news headlines were all accurate and selected from mainstream news outlets to be roughly contemporary with the false news headlines. Moreover, the headlines were selected to be either pro-Democratic or pro-Republican (and equally so). This was done using a pretest, which confirmed that the headlines were equally partisan across the categories (similar approaches have been described previously[11,19,20]). See Supplementary Information section 1 for details about the pretest.

Participants in study 1 were also asked: 'How important is it to you that you only share news articles on social media (such as Facebook and Twitter) if they are accurate?', to which they responded on a five-point scale from 'not at all important' to 'extremely important'. We also asked participants about their frequency of social media use, along with several exploratory questions about media trust. At the end of the survey, participants were asked whether they responded randomly at any point during the survey or searched for any of the headlines online (for example, via Google). As noted in our preregistration, we did not intend to exclude these individuals. Participants also completed several additional measures as part of separate investigations (this was also noted in the preregistration); namely, the seven-item cognitive reflection test[19], a political knowledge questionnaire, and the positive and negative affective schedule[35]. In addition, participants were asked several demographic questions (age, gender, education, income, and a variety of political and religious questions). The most central political partisanship question was 'Which of the following best describes your political preference?' followed by the following response options: strongly Democratic; Democratic; lean Democratic; lean Republican; Republican; and strongly Republican. For purposes of data analysis, this was converted to a Democratic or Republican binary variable. The full survey is available online in both text format and as a Qualtrics file, along with all data (https://osf.io/p6u8k/).

Procedure. Participants in the accuracy condition were given the following instructions: 'You will be presented with a series of news headlines from 2017 to 2019 (36 in total). We are interested in whether you think these headlines describe an event that actually happened in an accurate and unbiased way. Note: the images may take a moment to load.' In the sharing condition, the middle sentence was replaced with 'We are interested in whether you would consider sharing these stories on social media (such as Facebook or Twitter)'. We then presented participants with the full set of headlines in a random order. In the accuracy condition, participants were asked 'To the best of your knowledge, is this claim in the above headline accurate?' In the sharing condition, participants were asked 'Would you consider sharing this story online (for example, through Facebook or Twitter)?' Although these sharing decisions are hypothetical, headline-level analyses suggest that self-report sharing decisions of news articles such as those used in our study correlate strongly with actual sharing on social media[36].

In both conditions, the response options were simply 'no' and 'yes'. Moreover, participants saw the response options listed as either yes/no or no/yes (randomized across participants—that is, an individual participant only ever saw 'yes' first or 'no' first).

This study was approved by the University of Regina Research Ethics Board (Protocol 2018-116).

Analysis plan. Our preregistration specified that all analyses would be performed at the level of the individual item (that is, one data point per item per participant; 0 = no, 1 = yes) using linear regression with robust standard errors clustered on participant. However, we subsequently realized that we should also be clustering standard errors on headline (as multiple ratings of the same headline are non-independent in a similar way to multiple ratings from the same participant), and thus deviated from the preregistrations in this minor way (all key results are qualitatively equivalent if only clustering standard errors on participant). The linear regression was preregistered to have the following independent variables: a condition dummy (−0.5 = accuracy, 0.5 = sharing), a news type dummy (−0.5 = false, 0.5 = true), a political concordance dummy (−0.5 = discordant, 0.5 = concordant), and all two-way and three-way interactions. (Political concordance is defined based on the match between content and ideology; specifically, political concordant = pro-Democratic [pro-Republican] news (based on a pretest) for American individuals who prefer the Democratic [Republican] party over the Republican [Democratic]. Politically discordant is the opposite.) Our key prediction was that there would be a negative interaction between condition and news type, such that the difference between false and true is smaller in the sharing condition than the accuracy condition. A secondary prediction was that there would be a positive interaction between condition and concordance, such that the difference between concordant and discordant is larger in the sharing condition than the accuracy condition. We also said we would check for a three-way interaction, and use a Wald test of the relevant net coefficients to test how sharing likelihood of false concordant headlines compares to

true discordant headlines. Finally, as robustness checks, we said we would repeat the main analysis using logistic regression instead of linear regression, and using ratings that are *z*-scored within condition.

Study 2

Study 2 extended the observation of study 1 that most people self-report that it is important to not share accuracy information on social media. First, study 2 assesses the relative, in addition to absolute, importance placed on accuracy by also asking about the importance of various other factors. Second, study 2 tested whether the results of study 1 would generalize beyond MTurk by recruiting participants from Lucid for Academics, delivering a sample that matches the distribution of American residents on age, gender, ethnicity and geographical region. Third, study 2 avoided the potential spillover effects from study 1 condition assignment suggested in Extended Data Fig. 1 by not having participants complete a task related to social media beforehand.

In total, 401 participants (mean age of 43.7) completed the survey on 9–12 January 2020, including 209 males and 184 females, and 8 indicating other gender identities. Participants were asked 'When deciding whether to share a piece of content on social media, how important is it to you that the content is...' and then were given a response grid where the columns were labelled 'not at all', 'slightly', 'moderately', 'very', and 'extremely', and the rows were labelled 'accurate', 'surprising', 'interesting', 'aligned with your politics' and 'funny'.

This study was approved by the MIT COUHES (protocol 1806400195).

Studies 3, 4 and 5

In studies 3, 4, and 5 we investigate whether subtly shifting attention to accuracy increases the veracity of the news people are willing to share. In particular, participants were asked to judge the accuracy of a single (politically neutral) news headline at the beginning of the study, ostensibly as part of a pretest for another study. We then tested whether this accuracy-cue affects the tendency of individuals to discern between false and true news when making subsequent judgments about social media sharing. The principal advantage of this design is that the manipulation is subtle and not explicitly linked to the main task. Thus, although social desirability bias may lead people to underreport their likelihood of sharing misinformation overall, it is unlikely that any between-condition difference is driven by participants believing that the accuracy question at the beginning of the treatment condition was designed to make them take accuracy into account when making sharing decisions during the main experiment. It is therefore relatively unlikely that any treatment effect on sharing would be due to demand characteristics or social desirability.

The only difference between studies 3 and 4 was the set of headlines used, to demonstrate the generalizability of these findings. Study 5 used a more representative sample and included an active control condition and a second treatment condition that primed accuracy concerns in a different way. Studies 3 and 4 were approved by the Yale University Committee for the Use of Human Subjects (IRB protocol 1307012383). Study 5 was approved by the University of Regina Research Ethics Board (protocol 2018-116).

Participants. In study 3, we preregistered a target sample of 1,200 participants from MTurk. In total, 1,254 participants began the survey between 4–6 October 2017. However, 21 participants reporting not having a Facebook profile at the outset of the study and, as per our preregistration, were not allowed to proceed; and 71 participants did not complete the survey. The full sample (mean age of 33.7) included 453 males, 703 females, and 2 who did not answer the question. Following the main task, participants were asked whether they 'would ever consider sharing something political on Facebook' and were given the following response options: 'yes', 'no', and 'I don't use social media'. As per our preregistration, only participants who selected 'yes' to this question were included in our main analysis. This excluded 431 people and the sample of participants who would consider sharing political content (mean age of 34.5) included 274 males, 451 females, and 2 who did not answer the gender question.

In study 4, we preregistered a target sample of 1,200 participants from MTurk. In total, 1,328 participants began the survey between 28–30 November 2017. However, 8 participants did not report having a Facebook profile and 72 participants did not finish the survey. The full sample (mean age of 33.3) included 490 males, 757 females, and 1 who did not answer the question. Restricting to participants were responded 'Yes' when asked whether they 'would ever consider sharing something political on Facebook' excluded 468 people, such that the sample of participants who would consider sharing political content (mean age of 33.6) included 282 males, 497 females, and 1 who did not answer the gender question.

In study 5, we preregistered a target sample of 1,200 participants from Lucid. In total, 1,628 participants began the survey between 30 April and 1 May 2019. However, 236 participants reported not having a Facebook profile (and thus were not allowed to complete the survey) and 105 participants did not finish the survey. The full sample (mean age of 45.5) included 626 males and 661 females. Restricting to participants were responded 'yes' when asked whether they 'would ever consider sharing something political on Facebook' excluded 616 people, such that the sample of participants who would consider sharing political content (mean age of 44.3) included 333 males and 338 females.

Unlike in study 1, because the question about ever sharing political content was asked after the experimental manipulation (rather than at the outset of the study), there is the possibility that excluding participants who responded 'no' may introduce selection effects and undermine causal inference[37]. Although there was no significant difference in responses to this political sharing question between conditions in any of the three accuracy priming experiments (χ^2 test; study 3: $\chi^2(1, n = 1{,}158) = 0.156$, $P = 0.69$; study 4: $\chi^2(1, n = 1{,}248) = 0.988$, $P = 0.32$; study 5, $\chi^2(3, n = 1{,}287) = 2.320$, $P = 0.51$), for completeness we show that the results are robust to including all participants (see Supplementary Information section 2).

Materials. In study 3, we presented participants with 24 news headlines from ref. [20]; in study 4, we presented participants with a different set of 24 news headlines selected via pretest; and in study 5, we presented participants with yet another set of 20 news headlines selected via pretest. In all studies, half of the headlines were false (selected from a third-party fact-checking website, www.Snopes.com, and therefore verified as being fabricated and untrue) and the other half were true (accurate and selected from mainstream news outlets to be roughly contemporary with the false news headlines). Moreover, half of the headlines were pro-Democratic or anti-Republican and the other half were pro-Republican or anti-Democrat (as determined by the pretests). See Supplementary Information section 1 for further details on the pretests.

As in study 1, after the main task, participants in studies 3–5 were asked about the importance of sharing only accurate news articles on social media (study 4 also asked about the importance participants' friends placed on sharing only accurate news on social media). Participants then completed various exploratory measures and demographics. The demographics included the question 'If you absolutely had to choose between only the Democratic and Republican party, which would do you prefer?' followed by the following response options: Democratic Party or Republican Party. We use this question to classify participants as Democrats versus Republicans.

Procedure. In all three studies, participants were first asked whether they had a Facebook account, and those who did not were not permitted to complete the study. Participants were then randomly assigned to one of two conditions in studies 3 and 4, and one of four conditions in study 5.

In the 'treatment' condition of all three studies, participants were given the following instructions: 'First, we would like to pretest an actual news headline for future studies. We are interested in whether people think it is accurate or not. We only need you to give your opinion about the accuracy of a single headline. We will then continue on to the primary task. Note: the image may take a moment to load.' Participants were then shown a politically neutral headline and were asked: 'To the best of your knowledge, how accurate is the claim in the above headline?' and were given the following response scale: 'not at all accurate', 'not very accurate', 'somewhat accurate', 'very accurate'. One of two politically neutral headlines (1 true, 1 false) was randomly selected in studies 3 and 4; one of four politically neutral headlines (2 true, 2 false) was randomly selected in study 5.

In the 'active control' condition of study 5, participants were instead given the following instructions: 'First, we would like to pretest an actual news headline for future studies. We are interested in whether people think it is funny or not. We only need you to give your opinion about the funniness of a single headline. We will then continue on to the primary task. Note: the image may take a moment to load.' They were then presented with one of the same four neutral news headlines used in the treatment condition and asked: 'In your opinion, is the above headline funny, amusing, or entertaining?'. (Response options: extremely unfunny; moderately unfunny; slightly unfunny; slightly funny; moderately funny; extremely funny.)

In the 'importance treatment' condition of study 5, participants were instead asked the following question at the outset of the study: 'Do you agree or disagree that 'it is important to only share news content on social media that is accurate and unbiased'?'. (Response options: strongly agree to strongly disagree.)

In the 'control' condition of all three studies, participants received no initial instructions and proceeded directly to the next step.

Participants in all conditions were then told: 'You will be presented with a series of news headlines from 2016 and 2017 (24 in total) [2017 and 2018 (20 in total) for study 5]. We are interested in whether you would be willing to share the story on Facebook. Note: The images may take a moment to load.' They then proceeded to the main task in which they were presented with the true and false headlines and for each were asked 'If you were to see the above article on Facebook, how likely would you be to share it' and given the following response scale: 'extremely unlikely, moderately unlikely, slightly unlikely, slightly likely, moderately likely, extremely likely'. We used a continuous scale, instead of the binary scale used in study 1, to increase the sensitivity of the measure.

Analysis plan. Our preregistrations specified that all analyses would be performed at the level of the individual item (that is, one data point per item per participant, with the six-point sharing Likert scale rescaled to the interval [0, 1]) using linear regression with robust standard errors clustered on participant. However, we subsequently realized that we should also be clustering standard errors on headline (as multiple ratings of the same headline are non-independent in a similar way to multiple ratings from the same participant), and thus deviated from the preregistrations in this minor way (all key results are qualitatively equivalent if only clustering standard errors on participant).

In studies 3 and 4, the key preregistered test was an interaction between a condition dummy (0 = control, 1 = treatment) and a news veracity dummy (0 = false, 1 = true). This was to be followed-up by tests for simple effects of news veracity in each of the two conditions; and, specifically, the effect was predicted to be larger in the treatment condition. We also planned to test for simple effects of condition for each of the two types of news; and, specifically, the effect was predicted to be larger for false relative to true news. We also conducted a post hoc analysis using a linear regression with robust standard errors clustered on participant and headline to examine the potential moderating role of a dummy for the participant's partisanship (preference for the Democratic versus Republican party) and a dummy for the headline's political concordance (pro-Democratic [pro-Republican] headlines scored as concordant for participants who preferred the Democratic [Republican] party; pro-Republican [pro-Democratic] headlines scored as discordant for participants who preferred the Democratic [Republican] party). For ease of interpretation, we *z*-scored the partisanship and concordance dummies, and then included all possible interactions in the regression model. To maximize statistical power for these moderation analyses, we pooled the data from studies 3 and 4.

In study 5, the first preregistered test was to compare whether the active and passive control conditions differed, by testing for significant a main effect of condition (0 = passive, 1 = active), or significant interaction between condition and news veracity (0 = false, 1 = true). If these did not differ, we preregistered that we would combine the two control conditions for subsequent analyses. We would then test whether the two treatment conditions differ from the control condition(s) by testing for an interaction between dummies for each treatment (0 = passive or active control, 1 = treatment being tested) and news veracity. This was to be followed-up by tests for simple effects of news veracity in each of the conditions; and, specifically, the effect was predicted to be larger in the treatment conditions. We also planned to test for simple effects of condition for each of the two types of news; and, specifically, the effect was predicted to be larger for false relative to true news.

Study 6

Studies 3, 4 and 5 found that a subtle reminder of the concept of accuracy decreased sharing of false (but not true) news. In study 6, we instead use a full-attention treatment that directly forces participants to consider the accuracy of each headline before deciding whether to share it. This allows us to determine, within this particular context, the maximum effect that can be obtained by focusing attention on accuracy. Furthermore, using the accuracy ratings elicited in the full-attention treatment, we can determine what fraction of shared content was believed to be accurate versus inaccurate by the sharer. Together, these analyses allow us to infer the fraction of sharing of false content that is attributable to inattention, confusion about veracity, and purposeful sharing of falsehood.

This study was approved by the Yale University Committee for the Use of Human Subjects (IRB protocol 1307012383).

Participants. We combine two rounds of data collection on MTurk, the first of which had 218 participants begin the study on 11 August 2017, and the second of which had 542 participants begin the study on 24 August 2017, for a total of 760 participants. However, 14 participants did not report having a Facebook profile and 33 participants did not finish the survey. The full sample (mean age of 34.0) included 331 males, 376 females, and 4 who did not answer the question. Participants were asked whether they 'would ever consider sharing something political on Facebook' and were given the following response options: 'yes', 'no', 'I don't use social media'. Only participants who selected 'yes' to this question were included in our main analysis, as in our other studies (there was no significant difference in responses between conditions, $\chi^2_{(2)} = 1.07$, $P = 0.585$). This excluded 313 people and the final sample (mean age of 35.2) included 181 males, 213 females, and 4 who did not answer the gender question. For robustness, we also report analyses including all participants; see Extended Data Table 2.

Materials. We presented participants with the same 24 headlines used in study 3.

Procedure. Participants were first asked if they have a Facebook account and those who did not were not permitted to complete the study. Participants were then randomly assigned to one of two conditions. In the full-attention treatment condition, participants were given the following instructions: 'You will be presented with a series of news

headlines from 2016 and 2017 (24 in total). We are interested in two things: (i) Whether you think the headlines are accurate or not. (ii) Whether you would be willing to share the story on Facebook. Note: the images may take a moment to load.' In the control condition, participants were told: 'You will be presented with a series of news headlines from 2016 and 2017 (24 in total). We are interested in whether you would be willing to share the story on Facebook. Note: the images may take a moment to load.' Participants in both conditions were asked 'If you were to see the above article on Facebook, how likely would you be to share it' and given the following response scale: 'extremely unlikely', 'moderately unlikely', 'slightly unlikely', 'slightly likely', 'moderately likely', 'extremely likely'. Crucially, in the treatment condition, before being asked the social media sharing question, participants were asked: 'To the best of your knowledge, how accurate is the claim in the above headline?' and given the following response scale: 'not at all accurate', 'not very accurate', 'somewhat accurate', 'very accurate'.

Analysis. The goal of our analyses is to determine what fraction of the sharing of false headlines is attributable to confusion (incorrectly believing the headlines are accurate), inattention (forgetting to consider the accuracy of the headlines; as per the inattention-based account), and purposeful sharing of false content (as per the preference-based account). We can do so by using the sharing intentions in both conditions, and the accuracy judgments in the 'full-attention' treatment (no accuracy judgments were collected in the control). Because participants in the full-attention treatment are forced to consider the accuracy of each headline before deciding whether they would share it, inattention to accuracy is entirely eliminated in the full-attention treatment. Thus, the difference in sharing of false headlines between control and full-attention treatment indicates the fraction of sharing in control that was attributable to inattention. We can then use the accuracy judgments to determine how much of the sharing of false headlines in the full-attention treatment was attributable to confusion (indicated by the fraction of shared headlines that participants rated as accurate) versus purposeful sharing (indicated by the fraction of shared headlines that participants rated as inaccurate).

Concretely, we do the analysis as follows. First, we dichotomize responses, classifying sharing intentions of 'extremely unlikely', 'moderately unlikely', and 'slightly unlikely' as 'unlikely to share' and 'slightly likely', 'moderately likely', and 'extremely likely' as 'likely to share'; and classifying accuracy ratings of 'not at all accurate' and 'not very accurate' as 'not accurate' and 'somewhat accurate' and 'very accurate' as 'accurate'. Then we define the fraction of sharing of false content due to each factor as follows:

$$f_{\text{Inattention}} = \frac{F_{\text{cont}} - F_{\text{treat}}}{F_{\text{cont}}}$$

$$f_{\text{Confusion}} = \frac{N^{\text{acc}}_{\text{treat}}}{N_{\text{treat}}} \frac{F_{\text{treat}}}{F_{\text{cont}}}$$

$$f_{\text{Purposeful}} = \frac{N^{\text{inacc}}_{\text{treat}}}{N_{\text{treat}}} \frac{F_{\text{treat}}}{F_{\text{cont}}}$$

In which, F_{cont} denotes the fraction of false headlines shared in the control; F_{treat} denotes the fraction of false headlines shared in the treatment group; N_{treat} denotes the number of false headlines shared in the treatment group, $N^{\text{acc}}_{\text{treat}}$ denotes the number of false headlines shared and rated accurate in the treatment group, and $N^{\text{inacc}}_{\text{treat}}$ denotes the number of false headlines shared and rated inaccurate in the treatment group.

For an intuitive visualization of these expressions, see Fig. 2d.

To calculate confidence intervals on our estimates of the relative effect of inattention, confusion, and purposeful sharing, we use bootstrapping simulations. We create 10,000 bootstrap samples by sampling with replacement at the level of the subject. For each sample, we calculate the difference in fraction of sharing of false information explained by each of the three factors (that is, the three pairwise comparisons). We then determine a two-tailed P value for each comparison by doubling the fraction of samples in which the factor that explains less of the sharing in the actual data are found to explain more of the sharing.

Preregistration. Although we did complete a preregistration in connection with this experiment, we do not follow it here. The analyses we preregistered simply tested for an effect of the manipulation on sharing discernment, as in studies 3–5. After conducting the experiment, we realized that we could analyse the data in an alternative way to gain insight into the relevant effect of the three reasons for sharing misinformation described in this Article. It is these (post hoc) analyses that we focus on. Notably, Extended Data Table 2 shows that equivalent results are obtained when analysing the two samples separately (the first being a pilot for the pre-registered experiment, and the second being the pre-registered experiment), helping to address the post hoc nature of these analyses.

Study 7

In study 7, we set out to test whether the results of the survey experiments in studies 3–5 would generalize to real sharing decisions 'in the wild', and to misleading but not blatantly false news. Thus, we conducted a digital field experiment on Twitter in which we delivered the same intervention from the 'treatment' condition of the survey experiments to users who had previously shared links to unreliable news sites. We then examined the effect of receiving the intervention on the quality of the news that they subsequently shared. The experiment was approved by Yale University Committee of the Use of Human Subjects IRB protocol 2000022539 and MIT COUHES Protocol 1806393160. Although all analysis code is posted online, we did not publicly post the data owing to privacy concerns (even with de-identified data, it may be possible to identify many of the users in the dataset by matching their tweet histories with publicly available data from Twitter). Researchers interested in accessing the data are asked to contact the corresponding authors.

Study 7 is an aggregation of three different waves of data collection, the details of which are summarized in Extended Data Table 3. (These are all of the data that we collected, and the decision to conclude the data collection was made before running any of the analyses reported in this Article.)

Participants. The basic experimental design involved sending a private direct message to users asking them to rate the accuracy of a headline (as in the 'treatment' condition of the survey experiments). Twitter only allows direct messages to be sent from account X to account Y if account Y follows account X. Thus, our first task was to assemble a set of accounts with a substantial number of followers (who we could then send direct messages to). In particular, we needed followers who were likely to share misinformation. Our approach was as follows.

First, we created a list of tweets with links to one of two news sites that professional fact-checkers rated as extremely untrustworthy[27] but that are nonetheless fairly popular: www.Breitbart.com and www.infowars.com. We identified these tweets by (i) retrieving the timeline of the Breitbart Twitter account using the Twitter REST API (Infowars had been banned from Twitter when we were conducting our experiment and thus had no Twitter account), and (ii) searching for tweets that contain a link to the corresponding domain using the Twitter advanced search feature and collecting the tweet IDs either manually (wave 1) or via scraping (waves 2 and 3). Next, we used the Twitter API to retrieve lists of users who retweeted each of those tweets (we periodically fetched the list of 'retweeters' because the Twitter API only provides the last 100 users 'retweeters' of a given tweet). As shown in Extended Data Table 3, across the three waves this process yielded a potential participant list

of 136,379 total Twitter users with some history of retweeting links to misleading news sites.

Next, we created a series of accounts with innocuous names (for example, 'CookingBot'); we created new accounts for each experimental wave. Each of the users in the potential participant list was then randomly assigned to be followed by one of our accounts. We relied on the tendency of Twitter users to reciprocally follow-back to create our set of followers. Indeed, 8.3% of the users that were followed by one of our accounts chose to follow our account back. This yielded a total of 11,364 followers across the three waves. (After the completion of our experiments, Twitter has made it substantially harder to follow large numbers of accounts without getting suspended, which creates a challenge for using this approach in future work; a solution is to use the targeted advertising on Twitter to target adverts whose goal is the accruing of followers as the set of users one would like to have in one's subject pool.)

To determine eligibility and to allow blocked randomization, we then identified (i) users' political ideology using the algorithm from Barberá et al.[38]; (ii) the probability of them being a bot, using the bot-or-not algorithm[39]; (iii) the number of tweets to one of the 60 websites with fact-checker ratings that will form our quality measure; and (iv) the average fact-checker rating (quality score) across those tweets.

For waves 1 and 2, we excluded users who tweeted no links to any of the 60 sites in our list in the two weeks before the experiment; who could not be given an ideology score; who could not be given a bot score; or who had a bot score above 0.5 (in wave 1, we also excluded a small number of very high-frequency tweeters for whom we were unable to retrieve all relevant tweets due to the 3,200-tweet limit of the Twitter API). In wave 3, we took a different approach to avoiding bots, namely avoiding high-frequency tweeters. Specifically, we excluded participants who tweeted more than 30 links to one of the 60 sites in our list in the two weeks before the experiment, as well as excluding those who tweeted fewer than 5 links to one of the 60 sites (to avoid lack of signal). This resulted in a total of 5,379 unique Twitter users across the three waves. (Note that these exclusions were applied ex ante, and excluded users were not included in the experiment, rather than implementing post hoc exclusions.)

One might be concerned about systematic differences between the users we included in our experiments versus those who we followed but did not follow us back. To gain some insight into this question, we compared the characteristics of the 5,379 users in our experiment to a random sample of 10,000 users that we followed but did follow us back (sampled proportional to the number of users in each wave). For each user we retrieved number of followers, number of accounts followed, number of favourites, and number of tweets. We also estimated political ideology as per Barberá et al.[38], probability of being a bot[39], and age and gender using based on profile pictures using the *Face Plus Plus* algorithm[40–42]. Finally, we checked whether the account had been suspended or deleted. As shown in Extended Data Fig. 5, relative to users who did not follow us back, the users that took part in our experiment followed more accounts, had more followers, selected more favourite tweets, were more conservative, were older, and were more likely to be bots ($P < 0.001$ for all); and were also more likely to have had their accounts suspended or deleted ($P = 0.012$). These observations suggest that to the extent that our recruitment process induced selection, it is in a direction that works against the effectiveness of our treatment: the users in our experiment are likely to be less receptive to the intervention than users more generally, and therefore our effect size is likely to be an underestimate of the effect we would have observed in the full sample.

Materials and procedure. The treatment in study 7 was very similar to the survey experiments. Users were sent a direct message asking them to rate the accuracy of a single non-political headline (Fig. 4b). An advantage of our design is that this direct message is coming from an account that the user has themselves opted in to following, rather than from a totally unknown account. Furthermore, the direct message begins by saying 'Thanks for following me!' and sending such thank-you direct messages is a common practice on Twitter. These factors should substantially mitigate any possibility of the users feeling suspicious or that they are being surveilled by our account, and instead make the direct message appear more a typical interaction on Twitter.

We did not expect users to respond to our message. Instead, our intervention was based on the idea that merely reading the opening line ('How accurate is this headline?') would make the concept of accuracy more salient. Because we could not reliably observe whether (or when) users read the message (because many users' privacy settings prevent the sending of read-receipts), we performed intent-to-treat analyses that included all subjects and assumed that treatment began as soon as the message was sent. Furthermore, to avoid demand effects, users were not informed that the message was being sent as part of a research study, and the accounts from which we sent the messages had innocuous descriptions (such as 'Cooking Bot'). Not informing users about the study was essential for ecological validity, and we felt that the scientific and practical benefits justified this approach given that the potential harm to participants was minimal, and the tweet data were all publicly available. See Supplementary Information section 4 for more discussion on the ethics of digital field experimentation.

Because of the rate limits of direct message imposed by Twitter, we could only send direct message to roughly 20 users per account per day. Thus, we conducted each wave in a series of 24-h blocks in which a small subset of users was sent a direct message each day. All tweets and retweets posted by all users in the experiment were collected on each day of the experiment. All links in these tweets were extracted (including expanding shortened URLs). The dataset was thus composed of the subset of these links that linked to one of 60 sites whose trustworthiness had been rated by professional fact-checkers in previous work[27] (with the data entry for a given observation being the trust score of the linked site).

To allow for causal inference, we used a stepped-wedge (also called randomized roll-out) design in which users were randomly assigned to a treatment date. This allows us to analyse tweets made during each of the 24-h treatment blocks, comparing tweets from users who received the direct message at the start of a given block ('treated') to tweets from users who had not yet been sent a direct message ('control). Because the treatment date is randomly assigned, it can be inferred that any systematic difference revealed by this comparison was caused by the treatment. (Wave 2 also included a subset of users who were randomly assigned to never receive the direct message.) To improve the precision of our estimate, random assignment to treatment date was approximately balanced across bot accounts in all waves, and across political ideology, number of tweets to rated sites in the two weeks before the experiment, and average quality of those tweets across treatment dates in waves 2 and 3.

Because our treatment was delivered via the Twitter API, we were vulnerable to unpredictable changes to, and unstated rules of, the API. These gave rise to several deviations from our planned procedure. On day 2 of wave 1, fewer than planned direct messages were sent as our accounts were blocked part way through the day; and no direct messages were sent on day 3 of wave 1 (hence, that day is not included in the experimental dataset). On day 2 of wave 2, Twitter disabled the direct message feature of the API for the day, so we were unable to send the direct messages in an automated fashion as planned. Instead, all 370 direct messages sent on that day were sent manually over the course of several hours (rather than simultaneously). On day 3 of wave 2, the API was once again functional, but partway through sending the direct messages, the credentials for our accounts were revoked and no further direct messages were sent. As a result, only 184 of the planned 369 direct messages were sent on that day. Furthermore, because we did not randomize the order of users across stratification blocks, the users on day 3 who were not sent a direct message were systematically different from those who were sent a direct message. (As discussed in detail

below, we consider analyses that use an intent-to-treat approach for wave 2 day 3—treating the data as if all 369 direct messages had indeed been sent—as well as analyses that exclude the data from wave 2 day 3.)

Analysis plan

As the experimental design and the data were substantially more complex than the survey experiment studies and we lacked well-established models to follow, it was not straightforward to determine the optimal way to analyse the data in study 7. This is reflected, for example, in the fact that wave 1 was not preregistered, two different preregistrations were submitted for wave 2 (one before data collection and one following data collection but before analysing the data), and one preregistration was submitted for wave 3, and each of the preregistrations stipulated a different analysis plan. Moreover, after completing all three waves, we realized that all of the analyses proposed in the preregistrations do not actually yield valid causal inferences because of issues involving missing data (as discussed in more detail below in the 'Dependent variable' section). Therefore, instead of conducting a particular preregistered analysis, we consider the pattern of results across a range of reasonable analyses.

All analyses are conducted at the user–day level using linear regression with heteroscedasticity-robust standard errors clustered on user. All analyses include all users on a given day who have not yet received the direct message as well as users who received the direct message on that day (users who received the direct message more than 24 h before the given day are not included). All analyses use a post-treatment dummy (0 = user has not yet been sent a direct message, 1 = user received the direct message that day) as the key independent variable. We note that this is an intent-to-treat approach that assumes that all direct messages on a given day are sent at exactly the same time, and counts all tweets in the subsequent 24-h block as post-treatment. Thus, to the extent that technical issues caused tweets on a given day to be sent earlier or later than the specified time, this approach may underestimate the treatment effect.

The analyses we consider differ in the following ways: dependent variable, model specification, type of tweet considered, approach to handling randomization failure, and approach to determining statistical significance. We now discuss each of these dimensions in more detail.

1. Dependent variable. We consider three different ways of quantifying tweet quality. Across approaches, a key issue is how to deal with missing data. Specifically, on days when a given user does not tweet any links to rated sites, the quality of their tweeted links is undefined. The approach implied in our preregistrations was to simply omit missing user–days (or to conduct analyses at the level of the tweet). Because the treatment is expected to influence the probability of tweeting, however, omitting missing user–days has the potential to create selection and thus undermine causal inference (and tweet-level analyses are even more problematic). For example, if a user tweets as a result of being treated but would not have tweeted had they been in the control (or does not tweet as a result of treatment but would have tweeted have they been in the control), then omitting the missing user–days breaks the independence between treatment and potential outcomes ensured by random assignment. Given that only 47.0% of user-days contained at least one tweeted link to a rated site, such issues are potentially quite problematic. We therefore consider three approaches to tweet quality that avoid this missing data problem.

The first measure is the average relative quality score. This measure assigns each tweeted link a relative quality score by taking the previously described fact-checker trust rating[27] (quality score, [0, 1], available for 60 news sites) of the domain being linked to, and subtracting the baseline quality score (the average quality score of all pre-treatment tweets across all users in all of the experimental days). Each user–day is then assigned an average relative quality score by averaging the relative quality score of all tweets made by the user in question on the day in question; and users who did not tweet on a given day are assigned an average relative quality score of 0 (thus avoiding the missing data problem). Importantly, this measure is quite conservative because the (roughly half of) post-treatment user–days in which data are missing are scored as '0'. Thus, this measure assumes that the treatment had no effect on users who did not tweet on the treatment day. If, instead, non-tweeting users would have shown the same effect had they actually tweeted, the estimated effect size would be roughly twice as large as what we observed here. We note that this measure is equivalent to using average quality scores (rather than relative quality score) and imputing the baseline quality score to fill missing data (so assuming that on missing days, the user's behaviour matches the pre-treatment average).

The second measure is the summed relative quality score. This measure assigns each tweeted link a relative quality score in the same manner described above. A summed relative quality score of the user–day is then 0 plus the sum of the relative quality scores of each link tweeted by that user on that day. Thus, the summed relative quality score increases as a user tweets more and higher quality links, and decreases as the user tweets more and lower quality links; and, as for the average relative quality score, users who tweet no rated links received a score of 0. As this measure is unbounded in both the positive and negative directions, and the distribution contains extreme values in both directions, we winsorize summed relative quality scores by replacing values above the 95th percentile with the 95th percentile, and replacing values below the 5th percentile with values below the 5th percentile (our results are qualitatively robust to alternative choices of threshold at which to winsorize).

The third measure is discernment, or the difference in the number of links to mainstream sites versus misinformation sites shared on a given user–day. This measure is mostly closely analogous to the analytic approach taken in studies 2-4. To assess the effect of the intervention on discernment, we transform the data into long format such that there are two observations per user–day, one indicating the number of tweets to mainstream sites and the other indicating the number of tweets to misinformation sites (as previously defined[27]). We then include a source type dummy (0 = misinformation, 1 = mainstream) in the regression, and interact this dummy with each independent variable. The treatment increases discernment if there is a significant positive interaction between the post-treatment dummy and the source type dummy. As these count measures are unbounded in the positive direction, and the distributions contain extreme values, we winsorize by replacing values above the 95th percentile of all values with the 95th percentile of all values (our results are qualitatively robust to alternative choices of threshold at which to winsorize).

Finally, as a control analysis, we also consider the treatment effect on the number of tweets in each user–day that did not contain links to any of the 60 rated news sites. As this count measure is unbounded in the positive direction, and the distribution contains extreme values, we winsorize by replacing values above the 95th percentile of all values with the 95th percentile of all values (our results are qualitatively robust to alternative choices of threshold at which to winsorize).

2. Determining statistical significance. We consider the results of two different methods for computing *P* values for each model. The first is the standard approach, in which regression is used in conjunction with asymptotic inference using Huber–White cluster-robust sandwich standard errors clustered on user to calculate *P* values. The second uses Fisherian randomization inference (FRI) to compute an exact *P* value (that is, has no more than the nominal type I error rate) in finite samples[28,43–45]. FRI is non-parametric and thus does not require any modelling assumptions about potential outcomes. Instead, the stochastic assignment mechanism determined by redrawing the treatment schedule, exactly as done in the original experiment, determines the distribution of the test statistic under the null hypothesis[45]. On the basis of our stepped-wedge design, our treatment corresponds to the day on which the user receives the direct message. Thus, to perform

FRI, we create 10,000 permutations of the assigned treatment day for each user by re-running the random assignment procedure used in each wave, and recompute the *t*-statistic for the coefficient of interest in each model in each permutation. We then determine *P* values for each model by computing the fraction of permutations that yielded *t*-statistics with absolute value larger than the *t*-statistic observed in the actual data. Note that therefore, FRI takes into account the details of the randomization procedure that approximately balanced treatment date across bots in all waves, and across ideology, tweet frequency, and tweet quality in waves 2 and 3.

3. Model specification. We consider four different model specifications. The first includes wave dummies. The second post-stratifies on wave by interacting centred wave dummies with the post-treatment dummy. This specification also allows us to assess whether any observed treatment effect significantly differs across waves by performing a joint significance test on the interaction terms. The third includes date dummies. The fourth post-stratifies on date by interacting centred date dummies with the post-treatment dummy. (We note that the estimates produced by the first two specifications may be problematic if there are secular trends in quality and they are used in conjunction with linear regression rather than FRI, but we include them for completeness and because they are closest to the analyses we pre-registered; excluding them does not qualitatively change our conclusions.)

4. Tweet type. The analysis can include all tweets, or can focus only on cases in which the user retweets the tweet containing the link without adding any comment. The former approach is more inclusive, but may contain cases in which the user is not endorsing the shared link (for example, someone debunking an incorrect story may still link to the original story). Thus, the latter case might more clearly identify tweets that are uncritically sharing the link in question. More importantly, retweeting without comment (low-engagement sharing) exemplifies the kind of fast, low-attention action that is our focus (in which we argue that people share misinformation despite a desire to only share accurate information—because the attentional spotlight is focused on other content dimensions). Primary tweets are much more deliberate actions, ones in which it is more likely that the user did consider their action before posting (and thus where our accuracy nudge would be expected to be ineffective).

5. Article type. The analysis can include all links, or can exclude (as much as possible) links to opinion articles. Although the hyperpartisan and fake news sites in our list do not typically demarcate opinion pieces, nearly all of the mainstream sites include 'opinion' in the URL of opinion pieces. Thus, for our analyses that minimize opinion articles, we exclude the 3.5% of links (6.8% of links to mainstream sources) that contained '/opinion/' or '/opinions/' in the URL.

6. Approach to randomization failure. As described above, owing to issues with the Twitter API on day 3 of wave 2, there was a partial randomization failure on that day (many of the users assigned to treatment did not receive a direct message). We consider two different ways of dealing with this randomization failure. In the intent-to-treat approach, we include all users from the randomization-failure day (with the post-treatment dummy taking on the value 1 for all users who were assigned to be sent a direct message on that day, regardless of whether they actually received a direct message). In the exclusion approach, we instead drop all data from that day.

In the main text, we present the results of the specification in which we analyse retweets without comment, include links to both opinion and non-opinion articles, include wave fixed effects, calculate *P* values using FRI, and exclude data from the day on which a technical issue led to a randomization failure. Extended Data Table 4 presents the results of all specifications.

The primary tests of effects of the treatment compare differences in tweet quality for all eligible user–days. However, this includes many user–days for which there are no tweets to rated sites, which can occur, for example, because that user does not even log on to Twitter on that day. To quantify effect sizes on a more relevant subpopulation, we employ the principal stratification framework whereby each unit belongs to one of four latent type[29,30]: never-taker user–days (which would not have any rated tweets in either treatment or control), always-taker user–days (user–days where the user tweets rated links that day in both treatment and control), complier user–days (in which the treatment causes tweeting of rated links that day, which would not have occurred otherwise), and defier user–days (in which treatment prevents tweeting of rated links). Because the estimated treatment effects on whether a user tweets on a given day are mostly positive (although not statistically significant; see Supplementary Table 9), we assume the absence of defier user–days. Under this assumption, we can estimate the fraction of user–days that are not never-taker user–days (that is, are complier or always-taker user–days). This is then the only population on which treatment effects on rated tweet quality can occur, as the never-taker user–days are by definition unaffected by treatment with respect to rated tweets. We can then estimate treatment effects on quality and discernment on this possibly affected subpopulation by rescaling the estimates for the full population by dividing by the estimated fraction of non-never-taker user–days. These estimates are then larger in magnitude because they account for the dilution due to the presence of units that are not affected by treatment because they do not produce tweets whether in treatment or control.

Moreover, it is important to remember that our estimates of the effect size for our subtle, one-off treatment are conservative. Although our intent-to-treat approach necessarily assumes that the message was seen immediately—and thus counts all tweets in the 24 h after the message was sent as 'treated'—we cannot reliably tell when (or even if) any given user saw our message. Thus, it is likely that many of the tweets we are counting as post-treatment were not actually treated, and that we are underestimating the true treatment effect as a result.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Data and materials for studies 1 to 6 are available at https://osf.io/p6u8k/. Owing to privacy concerns, data from study 7 are available upon request.

Code availability

Code for all studies is available at https://osf.io/p6u8k/.

35. Watson, D., Clark, L. A. & Tellegen, A. Development and validation of brief measures of positive and negative affect: the PANAS scales. *J. Pers. Soc. Psychol.* **54**, 1063–1070 (1988).
36. Mosleh, M., Pennycook, G. & Rand, D. G. Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter. *PLoS One* **15**, e0228882 (2020).
37. Montgomery, J. M., Nyhan, B. & Torres, M. How conditioning on posttreatment variables can ruin your experiment and what to do about it. *Am. J. Pol. Sci.* **62**, 760–775 (2018).
38. Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A. & Bonneau, R. Tweeting from left to right: is online political communication more than an echo chamber? *Psychol. Sci.* **26**, 1531–1542 (2015).
39. Davis, C. A., Varol, O., Ferrara, E., Flammini, A. & Menczer, F. BotOrNot: a system to evaluate social bots. In *Proc. 25th International Conference Companion on World Wide Web* 273–274 (Association for Computing Machinery (ACM), 2016).
40. Chakraborty, A. et al. Who makes trends? Understanding demographic biases in crowdsourced recommendations. In *Proc. 11th Int. Conf. Web Soc. Media* 22–31 (ICWSM, 2017).
41. Kteily, N. S., Rocklage, M. D., McClanahan, K. & Ho, A. K. Political ideology shapes the amplification of the accomplishments of disadvantaged vs. advantaged group members. *Proc. Natl Acad. Sci. USA* **116**, 1559–1568 (2019).
42. An, J. & Weber, I. #greysanatomy vs. #yankees: Demographics and Hashtag Use on Twitter. In *Proc. 10th AAAI Conf. Web Soc. Media* 523–526 (ICWSM, 2016).

43. Rubin, D. B. Randomization analysis of experimental data: the fisher randomization test comment. *J. Am. Stat. Assoc.* **75**, 591 (1980).
44. Rosenbaum, P. R. *Observational Studies* 71–104 (Springer New York, 2002).
45. Imbens, G. W. & Rubin, D. B. *Causal Inference: For Statistics, Social, and Biomedical Sciences an Introduction* (Cambridge Univ. Press, 2015).

Acknowledgements We acknowledge feedback and comments from A. Bear, J. Jordan, D. Lazer, T. Rai and B. Mønsted, as well as funding from the Ethics and Governance of Artificial Intelligence Initiative of the Miami Foundation, the William and Flora Hewlett Foundation, the Omidyar Network, the John Templeton Foundation grant 61061, the Canadian Institutes of Health Research, and the Social Sciences and Humanities Research Council of Canada.

Author contributions G.P. and D.G.R. conceived of the research; G.P. and D.G.R. designed the survey experiments; A.A.A. and G.P. conducted the survey experiments; G.P. and D.G.R. analysed the survey experiments; Z.E., M.M. and D.G.R. designed the Twitter experiment; Z.E., M.M. and A.A.A. conducted the Twitter experiment; Z.E., M.M., D.E. and D.G.R. analysed the Twitter experiment; D.G.R. designed and analysed the limit-attention utility model; M.M. and D.G.R. designed and analysed the network simulations; G.P. and D.G.R. wrote the paper, with input from Z.E., M.M., A.A.A. and D.E. All authors approved the final manuscript.

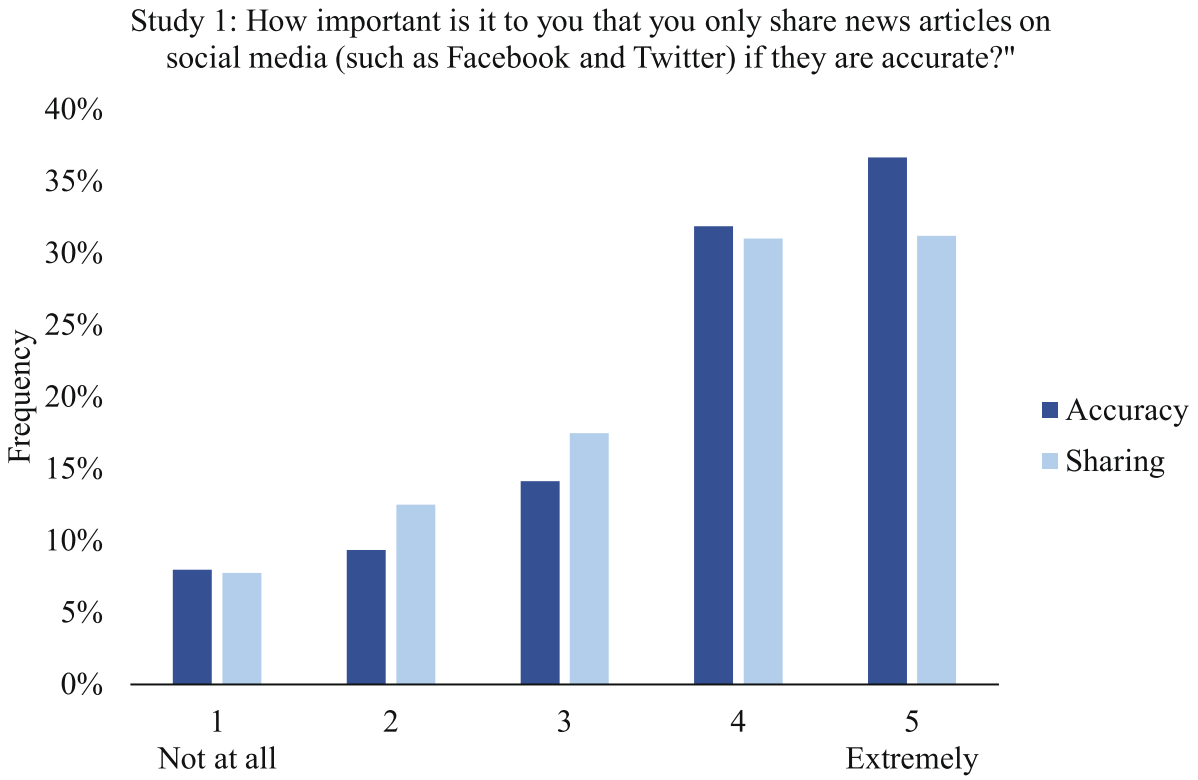
Competing interests Other research by D.E. is funded by Facebook, which has also sponsored a conference he co-organizes. D.E. previously had a significant financial interest in Facebook while contributing to this research. Other research by D.G.R. is funded by a gift from Google.

Additional information

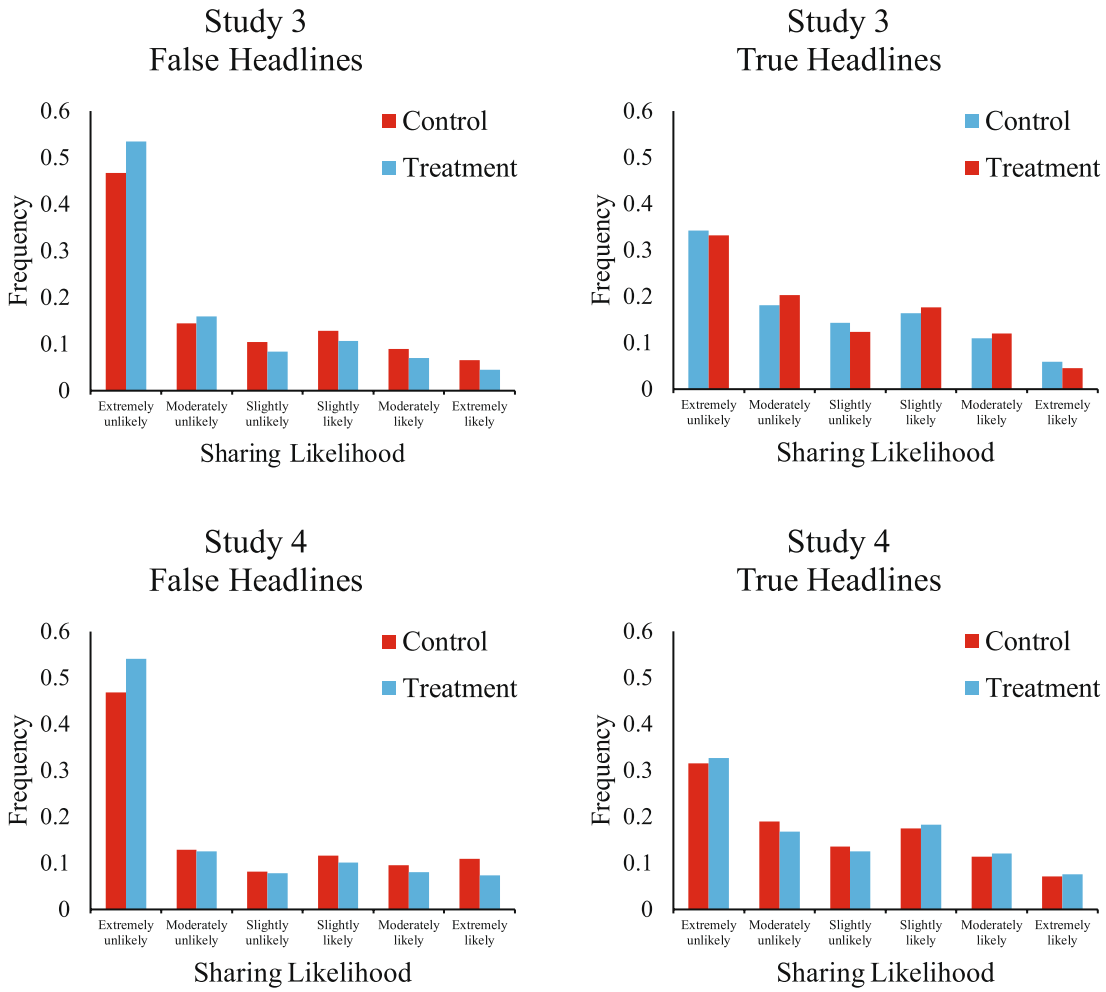
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03344-2.

Correspondence and requests for materials should be addressed to G.P. or D.G.R.

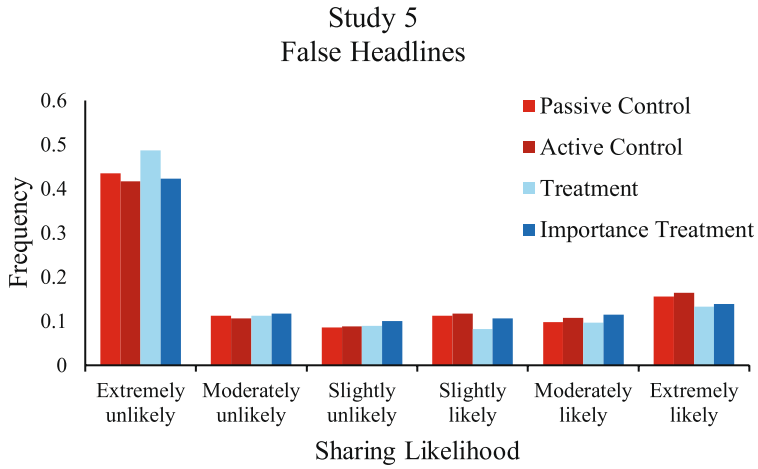
Peer review information *Nature* thanks David Lazer, Holly Fernandez Lynch and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

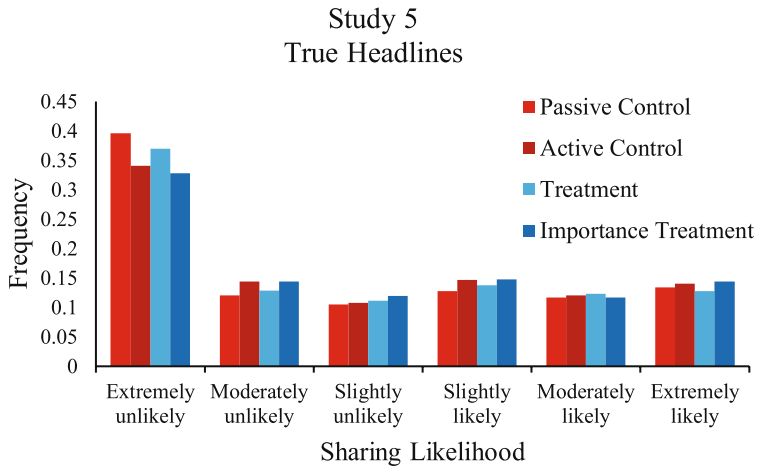


Extended Data Fig. 1 | Distribution of responses to the post-experimental question 'How important is it to you that you only share news articles on social media (such as Facebook and Twitter) if they are accurate' in study 1. Average responses were not statistically different in the sharing condition (mean = 3.65, s.d. = 1.25) compared to the accuracy condition (mean = 3.80, s.d. = 1.25) (*t*-test: $t_{(1003)} = 1.83$, $P = 0.067$).

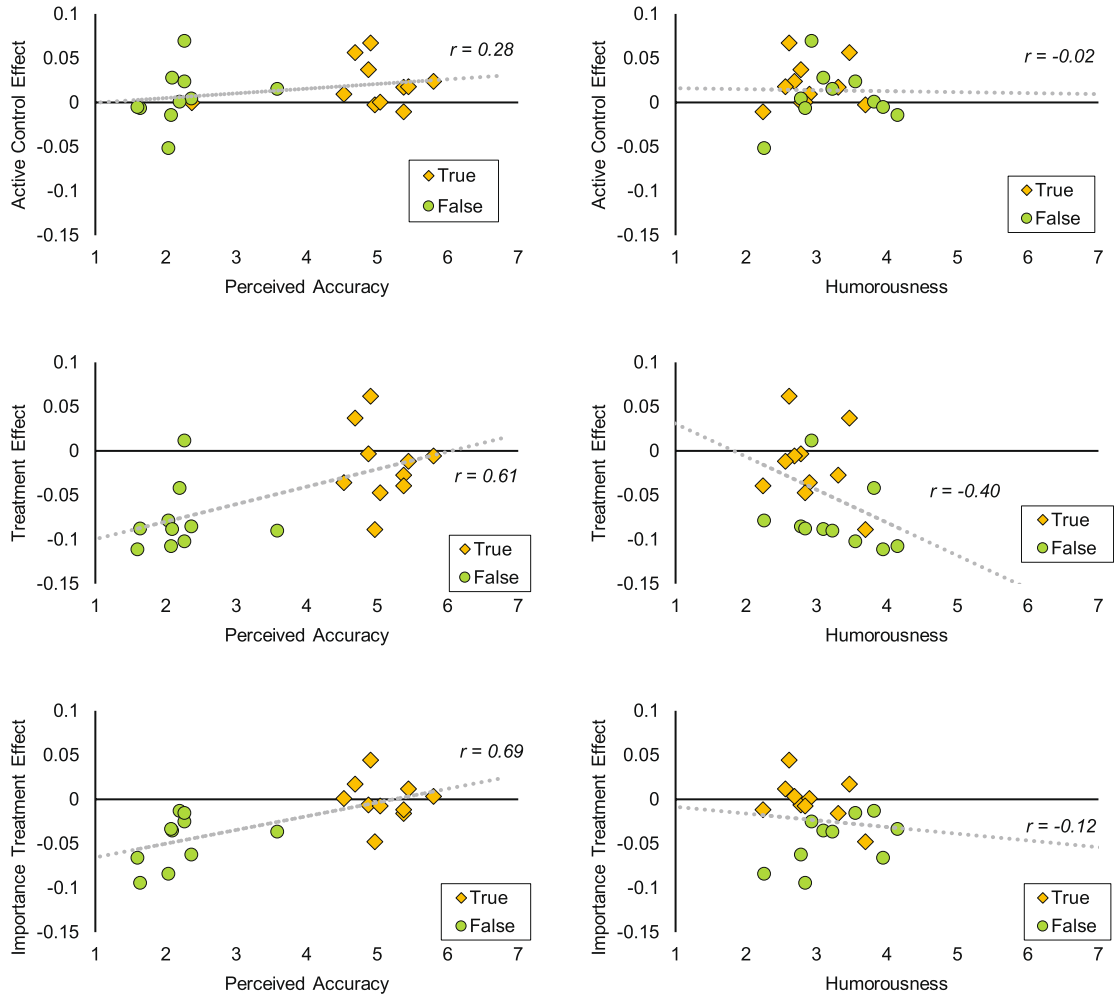


Extended Data Fig. 2 | Distribution of sharing intentions in studies 3 and 4, by condition and headline veracity. Whereas Fig. 2 discretizes the sharing intention variable for ease of interpretation such that all 'unlikely' responses are scored as 0 and all 'likely' responses are scored as 1, here the full distributions are shown. The regression models use these non-discretized values.

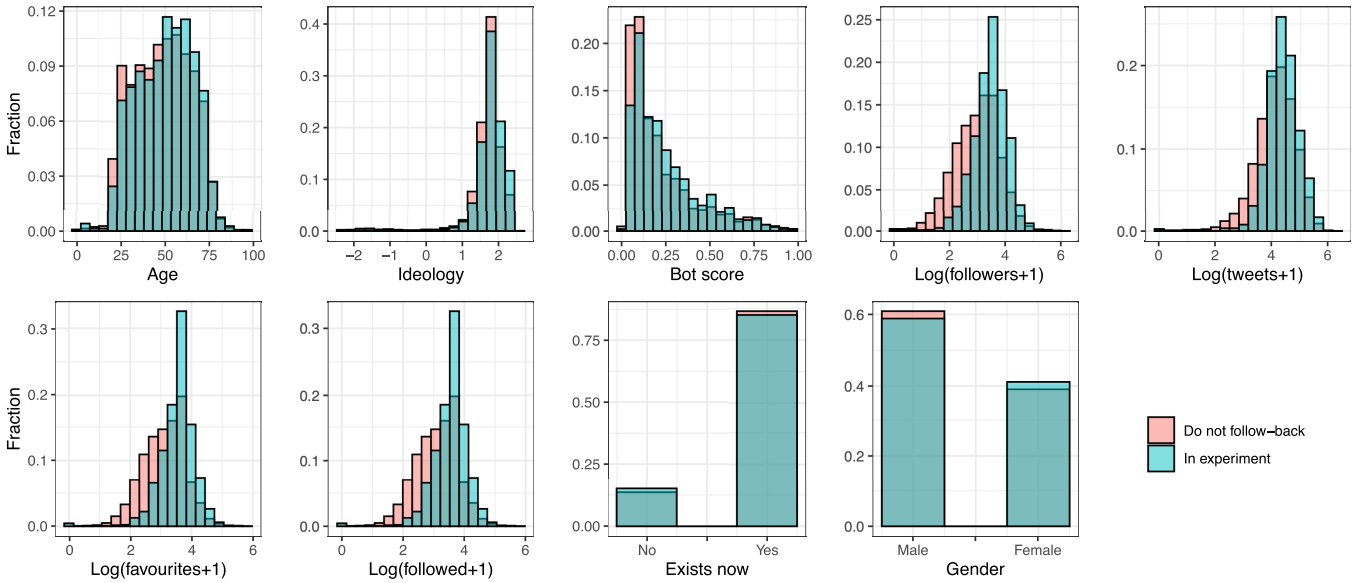




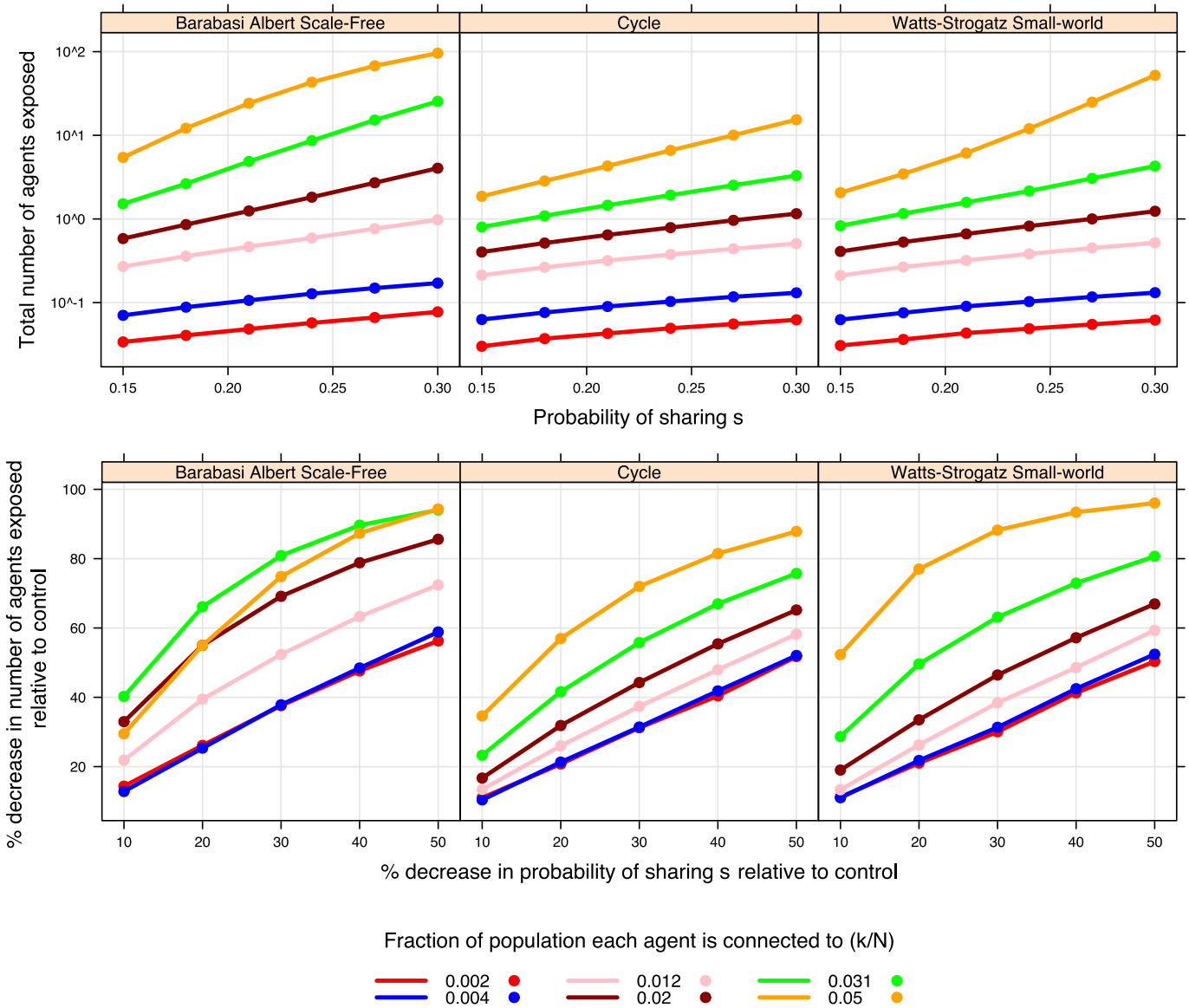
Extended Data Fig. 3 | Distribution of sharing intentions in study 5, by condition and headline veracity. Whereas Fig. 2 discretizes the sharing intention variable for ease of interpretation such that all 'unlikely' responses are scored as 0 and all 'likely' responses are scored as 1, here the full distributions are shown. The regression models use these non-discretized values.



Extended Data Fig. 4 | Headline-level analyses for study 5 showing the effect of each condition relative to control as a function of the perceived accuracy and humorousness of the headlines. For each headline, we calculate the effect size as the mean sharing intention in the condition in question minus the control (among users who indicate that they sometimes share political content); and we then plot this difference against the pre-test ratings of perceived accuracy and humorousness of the headline. The effect of both treatments is strongly correlated with the perceived accuracy of headline (treatment, $r_{(18)} = 0.61$, $P = 0.005$; importance treatment, $r_{(18)} = 0.69$, $P = 0.0008$), such that both treatments reduce sharing intentions to a greater extent as the headline becomes more inaccurate seeming. This supports our proposed mechanism in which the treatments operate through drawing attention to the concept of accuracy. Notably, we see no such analogous effect for the active control. Drawing attention to the concept of humorousness does not make people significantly less likely to share less humorous headlines (or more likely to share more humorous headlines), $r_{(18)} = -0.02$, $P = 0.93$. This confirms the prediction generated by our model fitting in Supplementary Information section 3.6—because our participants do not have a strong preference for sharing humorous news headlines, drawing their attention to humorousness does not influence their choices. This also demonstrates the importance of our theoretical approach that incorporates the role of preferences, relative to how priming is often conceptualized in psychology: drawing attention to a concept does not automatically lead to a greater effect of that concept on behaviour.



Extended Data Fig. 5 | Characteristics of the users in the study 7 Twitter field experiment. Users in the Twitter field experiment (blue) are compared to a random sample of 10,000 users who we followed but who did not follow-back our accounts (red). Relative to users who did not follow us back, the users who took part in our experiment followed more accounts, had more followers, selected more favourite tweets, were more conservative, were older and were more likely to be bots ($P < 0.001$ for all); and were also more likely to have had their accounts suspended or deleted ($P = 0.012$). These observations suggest that to the extent that our recruitment process induced selection, it is in a direction that works against the effectiveness of our treatment: the users in our experiment are likely to be less receptive to the intervention than users more generally, and therefore our effect size is likely to be an underestimate of the effect that we would have observed in the full sample.



Extended Data Fig. 6 | Results of agent-based simulations of news sharing on social networks. See Supplementary Information section 6 for model details. Shown is the relationship between individual-level probability of sharing misinformation and population-level exposure rates, for various levels of network density (fraction of the population that the average agent is connected to, *k*/*N*) and different network structures. Top, the raw number of agents exposed to the misinformation (*y* axis) as a function of the agents' raw probability of misinformation sharing (*x* axis). Bottom, the percentage reduction in the fraction of the population exposed to the piece of misinformation relative to control (*y* axis) as a function of the percentage reduction in individuals' probability of sharing the misinformation relative to control (*x* axis). As can be seen, a robust pattern emerges across network structures. First, we see that the network dynamics never suppress the individual-level intervention effect: a decrease in sharing probability of *x*% always decreases the fraction of the population exposed to the misinformation by at least *x*%. Second, in some cases the network dynamics can markedly amplify the effect of the individual-level intervention: for example, a 10% decrease in sharing probability can lead to up to a 40% decrease in the fraction of the population that is exposed, and a 50% decrease in sharing probability can lead to more than a 95% reduction in the fraction of the population that is exposed. These simulation results help to connect our findings about individual-level sharing to the resulting effects on population-level spreading dynamics of misinformation. They demonstrate the potential for individual-level interventions, such as the accuracy prompts that we propose here, to meaningfully improve the quality of the information that is spread via social media. These simulations also lay the groundwork for future theoretical work that can investigate a range of issues, including which agents to target if only a limited number of agents can be intervened on, the optimal spatiotemporal intervention schedule to minimize the frequency of any individual agent receiving the intervention (to minimize adaption or familiarity effects), and the inclusion of strategic sharing considerations (by introducing game theory).

Extended Data Table 1 | Best-fit parameter values and quantities of interest for the limited-attention utility model

	Study 4			Study 5		
Parameter	**Estimate**	**95% CI**		**Estimate**	**95% CI**	
β_P	0.35	0.25	0.51	1.22	0.97	1.45
β_H	-0.12	-0.21	0.12	0.57	0.40	0.87
p_{1c}	0.18	0.04	0.33	0.12	0.08	0.17
p_{2c}	0.22	0.09	0.47	0.48	0.42	0.52
p_{1t}	0.51	0.30	0.57	0.18	0.14	0.22
p_{2t}	0.00	0.00	0.34	0.51	0.46	0.55
θ	5.28	3.91	10.73	54.17	21.16	4091.50
k	-0.12	-0.20	-0.05	-0.03	-0.18	0.05
Overall probability considered in Control:						
Accuracy	0.40	0.33	0.59	0.60	0.54	0.65
Political Concordance	0.78	0.53	0.91	0.53	0.48	0.58
Humorousness	0.82	0.67	0.96	0.88	0.83	0.92
Overall probability considered in Treatment:						
Accuracy	0.51	0.46	0.64	0.68	0.63	0.73
Political Concordance	1.00	0.66	1.00	0.49	0.45	0.54
Humorousness	0.49	0.43	0.70	0.82	0.78	0.86
Treatment effect on probability of being considered:						
Accuracy	0.11	0.03	0.20	0.09	0.01	0.16
Political Concordance	0.22	0.07	0.35	-0.03	-0.10	0.03
Humorousness	-0.33	-0.48	-0.14	-0.06	-0.12	0.00

Results of fitting the model described in Supplementary Information section 3 to the experimental data from studies 4 and 5. The parameters β_P and β_H indicate preference for partisan alignment and humorousness, respectively, relative to accuracy; p_{1c}, p_{2c}, p_{1t} and p_{2t} indicate probabilities of attending to various pairs of preference terms in each condition (which are then used to construct the probabilities indicated lower in the table); and θ and k parameterize the sigmoid function that translates utility into choice. The key prediction of the preference-based account is that people care substantially less about accuracy than one or more of the other dimensions—that is, that $\beta_P > 1$ and/or $\beta_H > 1$. In contrast to this prediction, we see that β_H is significantly smaller than 1 in both studies (study 4, $P < 0.001$; study 5, $P = 0.001$), such that participants value accuracy more than humorousness; and β_P is significantly less than 1 in study 4 ($P < 0.001$), and not significantly different from 1 is study 5 ($P = 0.065$), such that participants value accuracy as much or more than political concordance. Thus, we find no evidence that participants care more about partisanship than accuracy. By contrast, this observation is consistent with the inattention-based account's prediction that participants value accuracy as much as, or more than, other dimensions. The results also confirm the inattention-based account's second prediction that by default (that is, in the control), participants often do not consider accuracy. Accordingly, we see that the probability of considering accuracy in the control is substantially lower than 1 (study 4, 0.40 [0.33, 0.59]; study 5, 0.60 [0.54, 0.65]). The confirmation of these two predictions provides quantitative support for the claim that inattention to accuracy has an important role in the sharing of misinformation in the control condition. Finally, the results confirm the inattention-based account's third prediction, namely that priming accuracy in the treatment will increase attention to accuracy; the probability that participants consider accuracy is significantly higher in the treatment compared to the control (study 4, $P = 0.005$; study 5, $P = 0.016$). *P* values calculated using bootstrapping.

Extended Data Table 2 | Fraction of sharing of false content attributable to inattention, confusion and purposeful sharing in study 6

	Political content sharers			All participants		
	Aggregate	Round 1	Round 2	Aggregate	Round 1	Round 2
Inattention	51.2%	53.7%	50.2%	50.8%	48.7%	51.6%
Confusion	33.1%	28.1%	35.0%	33.2%	31.3%	34.1%
Purposeful sharing	15.8%	18.2%	14.8%	16.0%	20.0%	14.3%

The results are extremely similar across rounds of data collection, and when including participants who do not report sharing political content online.

Extended Data Table 3 | Details for the three waves of study 7 data collection

Wave	Date Range	Treatment Time	Treatment Days	Bots	Users Followed	Follow-backs	Qualified Users	DMs sent	Link clicks	Rated tweets analyzed	Total tweets analyzed
1	4/20/2018-4/27/2018	7:43pm EST	7 (no 4/25)	6	19,913	821	705	705	80	12,912	231,162
2	9/12/2018-9/14/2018	5:00pm EST	3	7	23,673	3,111	2,153	1,060	60	24,912	387,993
3	1/28/2019-2/08/2019	7:00pm EST	12	13	92,793	7,432	2,521	2,330	169	15,918	564,843
Total			23	13	136,379	11,364	5,379	4,095	309	53,742	1,183,998

Extended Data Table 4 | Coefficients and *P* values associated with each model of quality for study 7

Tweet Type	Article Type	Randomization-Failure	Model Spec	Average Relative Quality			Summed Relative Quality			Discernment		
				Coeff	Reg p	FRI p	Coeff	Reg p	FRI p	Coeff	Reg p	FRI p
All	All	ITT	Wave FE	0.007	**0.004**	**0.009**	0.011	**0.022**	0.117	0.061	**0.004**	**0.016**
All	All	ITT	Wave PS	0.007	**0.006**	**0.009**	0.010	**0.020**	0.098	0.059	**0.004**	**0.018**
All	All	ITT	Date FE	0.006	**0.019**	**0.040**	0.009	0.070	0.267	0.053	**0.019**	0.055
All	All	ITT	Date PS	0.006	**0.041**	**0.035**	0.008	0.087	0.179	0.050	**0.028**	0.052
All	All	Exclude	Wave FE	0.007	**0.008**	**0.027**	0.013	**0.007**	0.074	0.065	**0.003**	**0.016**
All	All	Exclude	Wave PS	0.007	**0.011**	**0.024**	0.012	**0.009**	0.068	0.062	**0.003**	**0.019**
All	All	Exclude	Date FE	0.005	**0.045**	0.102	0.010	**0.044**	0.213	0.053	**0.020**	0.062
All	All	Exclude	Date PS	0.005	0.069	0.067	0.009	0.071	0.159	0.051	**0.032**	0.062
RT	All	ITT	Wave FE	0.007	**0.003**	**0.004**	0.012	**0.007**	**0.029**	0.058	**0.001**	**0.003**
RT	All	ITT	Wave PS	0.007	**0.004**	**0.004**	0.011	**0.006**	**0.020**	0.055	**0.001**	**0.003**
RT	All	ITT	Date FE	0.006	**0.017**	**0.014**	0.010	**0.032**	0.060	0.050	**0.008**	**0.006**
RT	All	ITT	Date PS	0.006	**0.027**	**0.012**	0.009	**0.042**	**0.035**	0.047	**0.016**	**0.013**
RT	All	Exclude	Wave FE	0.007	**0.004**	**0.009**	0.014	**0.002**	**0.011**	0.059	**0.001**	**0.003**
RT	All	Exclude	Wave PS	0.007	**0.005**	**0.008**	0.013	**0.002**	**0.011**	0.057	**0.001**	**0.004**
RT	All	Exclude	Date FE	0.006	**0.032**	**0.032**	0.011	**0.018**	**0.038**	0.049	**0.010**	**0.008**
RT	All	Exclude	Date PS	0.006	**0.042**	**0.023**	0.010	**0.033**	**0.027**	0.047	**0.021**	**0.017**
All	No Opinion	ITT	Wave FE	0.007	**0.002**	**0.015**	0.012	**0.012**	0.115	0.061	**0.004**	**0.017**
All	No Opinion	ITT	Wave PS	0.007	**0.004**	**0.016**	0.011	**0.011**	0.100	0.058	**0.004**	**0.021**
All	No Opinion	ITT	Date FE	0.006	**0.015**	0.057	0.010	0.051	0.271	0.054	**0.016**	**0.047**
All	No Opinion	ITT	Date PS	0.006	**0.031**	**0.044**	0.009	0.063	0.179	0.054	**0.018**	**0.034**
All	No Opinion	Exclude	Wave FE	0.007	**0.005**	**0.037**	0.014	**0.003**	0.067	0.064	**0.003**	**0.015**
All	No Opinion	Exclude	Wave PS	0.007	**0.008**	**0.035**	0.013	**0.005**	0.066	0.060	**0.003**	**0.019**
All	No Opinion	Exclude	Date FE	0.006	**0.033**	0.130	0.011	**0.027**	0.205	0.056	**0.015**	**0.047**
All	No Opinion	Exclude	Date PS	0.006	0.051	0.080	0.010	**0.047**	0.149	0.055	**0.019**	**0.036**
RT	No Opinion	ITT	Wave FE	0.008	**0.001**	**0.003**	0.012	**0.003**	**0.023**	0.057	**0.001**	**0.004**
RT	No Opinion	ITT	Wave PS	0.008	**0.002**	**0.004**	0.012	**0.003**	**0.019**	0.054	**0.001**	**0.004**
RT	No Opinion	ITT	Date FE	0.007	**0.009**	**0.013**	0.010	**0.022**	0.059	0.051	**0.006**	**0.007**
RT	No Opinion	ITT	Date PS	0.007	**0.013**	**0.007**	0.010	**0.026**	**0.028**	0.050	**0.010**	**0.008**
RT	No Opinion	Exclude	Wave FE	0.008	**0.001**	**0.009**	0.014	**0.001**	**0.008**	0.058	**0.001**	**0.004**
RT	No Opinion	Exclude	Wave PS	0.008	**0.003**	**0.008**	0.013	**0.001**	**0.009**	0.056	**0.001**	**0.005**
RT	No Opinion	Exclude	Date FE	0.006	**0.017**	**0.029**	0.011	**0.010**	**0.030**	0.051	**0.007**	**0.008**
RT	No Opinion	Exclude	Date PS	0.006	**0.021**	**0.014**	0.011	**0.018**	**0.019**	0.050	**0.013**	**0.011**

In the model specification column, 'FE' represents fixed effects (that is, just dummies) and 'PS' represents post-stratification (that is, centred dummies interacted with the post-treatment dummy). In the discernment column, the *P* value associated with the interaction between the post-treatment dummy and the source type dummy is reported; for all other dependent variables, the *P* value associated with the post-treatment dummy is reported. *P* values below 0.05 are in bold. Together, the results support the conclusion that the treatment significantly increased the quality of news shared. For the average relative quality score, virtually all (57 out of 64) analyses found a significant effect. For the summed relative quality score, most analyses found a significant effect, except for the FRI-derived *P* values when including all tweets. For discernment, 60 out of 64 analyses found a significant effect. Reassuringly, there was little qualitative difference between the two approaches for handling randomization failure, or across the four model specifications; and 98% of results were significant when only considering retweets without comment (which are the low-engagement sharing decisions that our theory predicts should respond to the treatment).

nature portfolio

https://doi.org/10.1038/s41586-021-03344-2

Supplementary information

Shifting attention to accuracy can reduce misinformation online

In the format provided by the authors and unedited

Supplementary Materials

for

Shifting attention to accuracy can reduce misinformation online

1. Pre-tests

Study 1

The pretest asked participants (N = 2,008 from MTurk, N = 1,988 from Lucid) to rate 10 randomly selected news headlines (from a corpus of 70 false, or 70 misleading/hyperpartisan, or 70 true) on a number of dimensions. Of primary interest, participants were asked the following question: "Assuming the above headline is entirely accurate, how favorable would it be to Democrats versus Republicans" – 1 = More favorable for Democrats, 5 = More favorable for Republicans). We used data from this question to selected the items used in Study 1 such that the Pro-Democratic items were equally different from the scale midpoint as the Pro-Republican items within the true and false categories. Participants were also asked to rate the headlines on the following dimensions: Plausibility ("What is the likelihood that the above headline is true" – 1 = Extremely unlikely, 7 = Extremely likely), Importance ("Assuming the headline is entirely accurate, how important would this news be?" - 1 = Extremely unimportant, 5 = Extremely important), Excitingness ("How exciting is this headline" - 1 = not at all, 5 = extremely), Worryingness ("How worrying is this headline?" - 1 = not at all, 5 = extremely), and Familiarity ("Are you familiar with the above headline (have you seen or heard about it before)?" – Yes/Unsure/No). Participants were also asked to indicate whether they would be willing to share each presented headline ("If you were to see the above article on social media, how likely would you be to share it?" - 1 = Extremely unlikely, 7 = Extremely likely). The pretest was run on June 24th, 2019.

Studies 3 and 6

For the pretest (completed on June 1st, 2017), participants (N = 209 from MTurk) rated 25 false headlines or 25 true headlines on the following dimensions: Plausibility ("What is the likelihood that the above headline is true" – 1 = Extremely unlikely, 7 = Extremely likely), Partisanship ("Assuming the above headline is entirely accurate, how favorable would it be to Democrats versus Republicans" – 1 = More favorable for Democrats, 5 = More favorable for Republicans), and Familiarity ("Are you familiar with the above headline (have you seen or heard about it before)?" -Yes/ Unsure/ No).

Study 4

For the pretest (completed on November 22nd, 2017), participants (N = 269 from MTurk) rated 36 false headlines or 36 true headlines on the following dimensions: Plausibility ("What is the likelihood that the above headline is true" – 1 = Extremely unlikely, 7 = Extremely likely), Partisanship ("Assuming the above headline is entirely accurate, how favorable would it be to Democrats versus Republicans" – 1 = More favorable for Democrats, 5 = More favorable for Republicans), Familiarity ("Are you familiar with the above headline (have you seen or heard about it before)?" -Yes/Unsure/No), and Humorousness ("In your opinion, is the above headline funny, amusing, or entertaining" 1 = extremely unfunny, 7 = extremely funny).

Study 5

The pretest asked participants (*N* = 516 from MTurk) to rate a random subset of 30 headlines from a larger set of false, hyperpartisan, and true headlines (there were 40 headlines in total in each category) on the following dimensions: Plausibility ("What is the likelihood that the above headline is true" – 1 = Extremely unlikely, 7 = Extremely likely), Partisanship ("Assuming the above headline is entirely accurate, how favorable would it be to Democrats versus Republicans" – 1 = More favorable for Democrats, 5 = More favorable for Republicans), Familiarity ("Are you familiar with the above headline (have you seen or heard about it before)?" – Yes/Unsure/No), Funniness ("In your opinion, is the above headline funny, amusing, or entertaining" 1 = extremely unfunny, 7 = extremely funny). The pretest was completed on May 23rd, 2018.

2. Regression tables

The full regression models are shown for Study 1 analyses in Table S1, for Studies 3 and 4 in Tables S2 and S3, and for Study 5 in Tables S4 and S5.

	(1) Linear Rating	(2) Logistic Rating	(3) Linear *z*-Rating
Condition (Accuracy=-0.5, Sharing=0.5)	-0.109***	-0.381***	-0.000407
	(0.0181)	(0.102)	(0.0377)
	1.65e-09	0.000186	0.991
Veracity (False=-0.5, True=0.5)	0.309***	1.460***	0.627***
	(0.0204)	(0.109)	(0.0422)
	<1e-10	<1e-10	<1e-10
Concordance of headline (-0.5=discordant, 0.5=concordant)	0.147***	0.741***	0.308***
	(0.0180)	(0.0992)	(0.0376)
	<1e-10	<1e-10	<1e-10
Condition X Veracity	-0.500***	-2.394***	-1.001***
	(0.0310)	(0.181)	(0.0637)
	<1e-10	<1e-10	<1e-10
Condition X Concordance	0.0917***	0.317**	0.208***
	(0.0221)	(0.115)	(0.0462)
	3.31e-05	0.00569	6.97e-06
Veracity X Concordance	0.0766*	0.252	0.159*
	(0.0348)	(0.191)	(0.0723)
	0.0274	0.188	0.0283
Condition X Veracity X Concordance	-0.0207	-0.0394	-0.0340
	(0.0396)	(0.203)	(0.0827)
	0.601	0.846	0.681
Constant	0.379***	-0.583***	0.000203
	(0.0113)	(0.0596)	(0.0234)
	<1e-10	<1e-10	0.993
Observations	36,180	36,180	36,180
Participant clusters	1005	1005	1005
Headline clusters	36	36	36
R-squared	0.207		0.189
Standard errors in parentheses; p-values below standard errors			
*** $p<0.001$, ** $p<0.01$, * $p<0.05$			

Table S1. Regressions with robust standard errors clustered on participant and headline predicting responses (0 or 1) in Study 1. Models 1 and 3 use linear regression; Model 2 uses logistic regression. Models 1 and 2 use the raw responses; Model 3 uses responses that are z-scored within condition. We observe a significant main effect of condition in Models 1 and 2, such that overall, participants were more likely to rate headlines as true than to say they would consider sharing them (this difference is eliminated by design in Model 3 because responses are z-scored within condition). Across all 3 models, we unsurprisingly observe significant positive main effects of veracity and concordance ($p < .001$ for both main effects in all models). Critically, as predicted, across all models we observe a significant negative interaction between condition and veracity, and a significant positive interaction between condition and headline concordance ($p < .001$ for both interactions in all models). Thus, participants are less sensitive to veracity, and more sensitive to concordance, when making sharing decisions than accuracy judgments. We also observe no significant 3-way interaction ($p > .100$ in all models). Finally, we see inconsistent evidence regarding a positive interaction between veracity and concordance, such that veracity may or may not play a bigger role among concordant headlines than discordant headlines.

	(1)	(2)	(3)	(4)	(5)	(6)
	Participants that share political content			All participants		
	S3	S4	S3+S4	S3	S4	S3+S4
Treatment	-0.0545***	-0.0582***	-0.0557***	-0.0294*	-0.0457***	-0.0372***
	(0.0145)	(0.0168)	(0.0110)	(0.0117)	(0.0139)	(0.00902)
	0.000176	0.000536	3.71e-07	0.0117	0.000977	3.79e-05
Veracity (0=False, 1=True)	0.0540**	0.0455	0.0494**	0.0383*	0.0378	0.0380**
	(0.0205)	(0.0271)	(0.0161)	(0.0169)	(0.0225)	(0.0138)
	0.00832	0.0934	0.00212	0.0237	0.0935	0.00590
Treatment X Veracity	0.0529***	0.0648***	0.0589***	0.0475***	0.0635***	0.0557***
	(0.0108)	(0.0147)	(0.00857)	(0.00818)	(0.0117)	(0.00681)
	8.74e-07	9.97e-06	<1e-10	6.69e-09	5.12e-08	<1e-10
z-Party (Prefer Republicans to Democrats)			0.0169			0.00902
			(0.00939)			(0.00804)
			0.0722			0.262
Veracity X Party			0.00322			0.00249
			(0.00930)			(0.00792)
			0.729			0.753
Treatment X Party			0.00508			0.0111
			(0.0106)			(0.00809)
			0.632			0.170
Treatment X Veracity X Party			-0.0159			-0.0113*
			(0.00864)			(0.00573)
			0.0663			0.0495
z-Concordance of Headline			0.0684***			0.0524***
			(0.00723)			(0.00625)
			<1e-10			<1e-10
Veracity X Concordance			0.00351			0.00396
			(0.0107)			(0.00897)
			0.743			0.659
Treatment X Concordance			-0.0156***			-0.00723*
			(0.00462)			(0.00315)
			0.000760			0.0219
Treatment X Veracity X Concordance			0.0224***			0.0163***
			(0.00527)			(0.00290)
			2.12e-05			1.90e-08
Party X Concordance			-0.00352			-0.00547
			(0.00928)			(0.00834)
			0.704			0.511
Treatment X Party X Concordance			0.00725			0.00930*
			(0.00471)			(0.00440)

			0.124			0.0347
Veracity X Party X Concordance			0.0157			0.0159
			(0.0135)			(0.0123)
			0.244			0.194
Treatment X Veracity X Party X Concordance			-0.0136**			-0.0132***
			(0.00448)			(0.00382)
			0.00241			0.000562
Constant	0.285***	0.314***	0.300***	0.234***	0.263***	0.249***
	(0.0152)	(0.0221)	(0.0125)	(0.0128)	(0.0182)	(0.0106)
	<1e-10	<1e-10	<1e-10	<1e-10	<1e-10	<1e-10
Observations	17,417	18,677	36,094	27,732	29,885	57,617
	727	780	1,507	1,158	1,248	2,406
	24	24	48	24	24	48
R-squared	0.019	0.016	0.063	0.012	0.014	0.045
Standard errors in parentheses; p-values below standard errors						
*** p<0.001, ** p<0.01, * p<0.05						

Table S2. Linear regressions predicting sharing intentions (1-6 Likert scale rescaled to [0,1]) in Studies 3 and 4. Robust standard errors clustered on participant and headline. In all cases, we observe (i) the predicted significant positive interaction between treatment and news veracity, such that sharing discernment was higher in the Treatment compared to the Control; (ii) a negative simple effect of condition for false headlines, such that participants were less likely to consider sharing false headlines in the Treatment compared to the Control; and (iii) no significant simple effect of condition for true headlines, such that participants were no less likely to consider sharing true headlines in the Treatment compared to the Control. Turning to potential moderation effects, we examine the regression models in columns 3 and 6. We see that the Treatment has a significantly larger effect on sharing discernment for concordant headlines (significant positive 3-way Treatment × Veracity × Concordance interaction); but that this moderation effect is driven by Democrats more so than Republicans (significant negative 4-way Treatment × Veracity × Concordance × Party interaction).

Simple effect	Net coefficient	Participants that share political content		All participants	
		S3	S4	S3	S4
Treatment on false headlines	Treatment	0.0002	0.0005	0.0117	0.0010
Treatment on true headlines	Treatment+Treatment×Veracity	0.9185	0.6280	0.1535	0.1149
Veracity in Control	Veracity	0.0083	0.0934	0.0237	0.0935
Veracity in Treatment	Veracity+Treatment×Veracity	<.0001	0.0001	<.0001	<.0001

Table S3. P-values associated with the various simple effects from the regression models in Table S2. Despite the significant interactions with concordance and partisanship, sharing of false headlines was significantly lower in the Treatment than the Control for every combination of participant partisanship and headline concordance ($p < .05$ for all), with the exception of Republicans sharing concordant headlines when including all participants ($p = .36$).

	(1)	(2)	(3)	(4)
	Participants that share political content		All participants	
	Controls only	All conditions	Controls only	All conditions
Veracity (0=False, 1=True)	0.00812	0.0163	0.0111	0.0154
	(0.0262)	(0.0234)	(0.0206)	(0.0212)
	0.756	0.486	0.589	0.466
Active Control	0.00606		0.0179	
	(0.0303)		(0.0223)	
	0.841		0.421	
Active Control X Veracity	0.0155		0.00856	
	(0.0120)		(0.00660)	
	0.199		0.195	
Treatment		-0.0815**		-0.0500**
		(0.0261)		(0.0185)
		0.00178		0.00685
Treatment X Veracity		0.0542***		0.0466***
		(0.0157)		(0.00914)
		0.000538		3.31e-07
Importance Treatment		-0.0504		-0.00966
		(0.0274)		(0.0193)
		0.0660		0.617
Importance Treatment X Veracity		0.0376**		0.0291***
		(0.0120)		(0.00634)
		0.00178		4.39e-06
Constant	0.477***	0.480***	0.359***	0.368***
	(0.0227)	(0.0160)	(0.0166)	(0.0127)
	<1e-10	<1e-10	<1e-10	<1e-10
Observations	6,776	13,340	12,847	25,587
Participant clusters	341	671	646	1286
Headline clusters	20	20	20	20
R-squared	0.001	0.007	0.001	0.004

Standard errors in parentheses; p-values below standard errors
*** p<0.001, ** p<0.01, * p<0.05

Table S4. Linear regressions predicting sharing intentions (1-6 Likert scale rescaled to [0,1]) in Study 5. Robust standard errors clustered on participant and headline. When comparing the passive and active controls, we see no significant main effect of condition or interaction with veracity, whether considering only participants who indicated that they sometimes consider sharing political content (Col 1) or all participants (Col 3). Therefore, as per our preregistered analysis plan, we collapse across control conditions for our main analysis. When comparing our main Treatment to the collapsed controls, we observed the predicted significant positive interaction between Treatment and news veracity, such that sharing discernment was higher in the Treatment compared to the controls, whether considering only participants who indicated that they sometimes consider sharing political content (Col 2) or considering all participants (Col 4). (Equivalent results are observed if comparing the Treatment only to the Active control.) When comparing our alternative Importance Treatment to the collapsed controls, we observed the predicted significant positive interaction between Importance Treatment and news veracity, such that sharing discernment was higher in the Importance Treatment compared to the controls, whether considering only participants who indicated that they sometimes consider sharing political content (Col 2) or considering all participants (Col 4).

Simple effect	Net coefficient	Participants that share political content	All participants
Treatment on false headlines	Treatment	0.0018	0.0068
Treatment on true headlines	Treatment+Treatment×Veracity	0.2411	0.8473
Importance Treatment on false headlines	Importance Treatment	0.0660	0.6166
Importance Treatment on true headlines	ImportanceTreatment+ImportanceTreatment×Veracity	0.5883	0.2700
Veracity in Controls	Veracity	0.4860	0.4665
Veracity in Treatment	Veracity+Treatment×Veracity	0.0032	0.0027
Veracity in Importance Treatment	Veracity+ImportanceTreatment×Veracity	0.0242	0.0470

Table S5. P-values associated with the various simple effects from the regression models in Table S4. We observe the predicted significant negative simple effect of Treatment for false headlines, such that participants were less likely to consider sharing false headlines in the Treatment compared to the controls; and no significant simple effect of Treatment for true headlines, such that participants were no less likely to consider sharing true headlines in the Treatment compared to the controls. The negative simple effect of the Importance Treatment for false headlines was only marginally significant when considering sharer participants and non-significant when considering all participants, and the simple effect of Importance Treatment for true headlines was non-significant in both cases. Thus the results for the Importance Treatment are somewhat weaker than for the main Treatment.

3. Formal model of social media sharing based on limited attention and preferences

Here we present a formal model to clearly articulate the competing hypotheses that we are examining. We then use this model to demonstrate the effectiveness of our experimental approach. Finally, we fit the model to our data in order to quantitatively support our inattention-based account of misinformation sharing.

The modeling framework we develop here combines three lines of theory. The first is utility theory, which is the cornerstone of economic models of choice[1–4]. When people are choosing across a set of options (in our case, whether or not to share a given piece of content), they preferentially choose the option which gives them more utility, and the utility they gain for a given choice is defined by their *preferences*. In virtually all such models, preferences are assumed to be fixed (or at least to change over much longer timescales than that of any specific decision, e.g. months or years). The second line of theorizing involves importance of attention. A core tenet of psychological theory is that when attention is drawn to a particular dimension of the environment (broadly construed), that dimension tends to receive more weight in subsequent decisions[5–8]. While attention has been a primary focus in psychology, it has only recently begun to be integrated with utility theory models – such that attention can increase the weight put on certain preferences over others when making decisions[9,10]. Another major body of work documents how our cognitive capacities are limited (and our rationality is bounded) such that we are not able to bring all relevant pieces of information to bear on a given decision[11–17]. While the integration of cognitive constraints and utility theory is a core topic in behavioral economics, this approach has typically not been applied to attention and the implementation of preferences. Thus, we develop a model in which attention operates via cognitive constraints: agents are limited to only considering a subset of their preferences in any given decision, and attention determines which preferences are considered.

3.1. Basic modeling framework

Consider a piece of content x which is defined by k different characteristic dimensions; one of these dimensions is whether the content is false/misleading $F(x)$, and the other k-1 dimensions are non-accuracy-related (e.g. partisan alignment, humorousness, etc) defined as $C_2(x)\ldots C_k(x)$. In our model, the utility a given person expects to derive from sharing content x is given by

$$U(x) = -a_1\beta_F F(x) + \sum_{i=2}^{k} a_i\beta_i C_i(x)$$

where β_F indicates how much they dislike sharing misleading content and $\beta_2 \ldots \beta_k$ indicate how much they care about each of the other dimensions (i.e. βs indicate preferences); while a_1 indicates how much the person is paying attention to accuracy, and $a_2\ldots a_k$ indicate how much the person is paying attention to each of the other dimensions. The probability that the person chooses to share the piece of content x is then increasing in $U(x)$. In the simplest decision rule, they will share if and only if $U(x) > 0$; for a more realistic decision rule, one could use the logistic function, such that

$$p(Share) = \frac{1}{1 + e^{-\theta(U(x)+k)}}$$

where k determines the value of $U(x)$ at which the person is equally likely to share versus not share, and θ determines the steepness of the transition around that point from sharing to not sharing (the simple decision rule described in the previous sentence corresponds to $k=0$, $\theta \rightarrow$ Inf).

In the standard utility theory model, $a_i=1$ for all i (all preferences are considered in every decision). In prior work on attention and preferences, a values are continuous, and are determined by some feature of the choice – for example, in the context of economic decisions, the difference between minimum and maximum possible payoffs [10], or the difference in percentage terms from the payoffs of other available lotteries [9]. Thus, all features are considered, but to differing degrees depending on how attention is focused.

In our limited-attention account, conversely, we incorporate cognitive constraints: we stipulate that people can consider only a subset of characteristic dimensions when making decisions. Specifically, agents can only attend to m out of the k utility terms in a given decision. That is, each value of a is either 0 or 1, $a_i \in \{0,1\}$; and because only m terms can be considered at once, the a values must sum to k, $\sum_{i=1}^{k} a_i = m$. Critically, the probability that any specific set of preference terms is attended to (i.e. which a values are equal to 1) is heavily influenced by the situation, and (unlike preferences) can change from moment to moment – in response, for example, to the application of a prime. As described below in Section 3.7, we provide evidence that our limited-attention formulation fits the experimental data better than the framework used in prior models of attention and preferences where all preferences are considered but with differing weights (despite our model having an equal number of free parameters). It is also important to note that our basic formulation takes attention (i.e. the probability that a given set of a_i values equal 1) as exogenously determined (e.g. by the context). However, in Section 3.7 we show that the results are virtually identical when using a more complex formulation where attention is also influenced by preferences, such that a person is more likely to pay attention to dimensions that they care more about (i.e. that have larger β values).

3.2. Preference-based versus inattention-based accounts

Within this framework, we can articulate the preference-based versus inattention-based accounts. The preference-based account stipulates that people care less about accuracy than other factors when deciding what to share. This idea reflects that argument that many people have a low regard for the truth when deciding what to share on social media (e.g., Lewandowsky, Ecker, & Cook, 2017). In terms of our model, this translates into the hypothesis that β_F is small compared to one or more of the other β terms – such that veracity has little impact on what content people decide to share (regardless of whether they are paying attention to it or not). Note that if the β values on accuracy and political concordance, for example, were equal, then people would only be likely to share content that they judged to be *both* accurate and politically concordant. The preference-based sharing of false, politically concordant content thus requires a substantially higher β on political concordance than on accuracy.

Our inattention-based account, conversely, builds off the contention that people often consider only a subset of characteristic dimensions when making decisions. Thus, even if people do have a strong preference for accuracy (i.e. β_F is as large, or larger than, other β values), how accurate content is may still have little impact on what people decide to share if the context focuses their limited attention on other dimensions. The accuracy-based account of misinformation sharing, then, is the hypothesis that (i) β_F is not appreciably smaller than the other β values (e.g. the β for political concordance), but that people nonetheless sometimes share misinformation because (ii) the probability of observing a_1=1 is far less than 1 ($p(a_1=1) << 1$), such that people often fail to consider accuracy. As a result, the inattention-based account (but not the preference-based account) predicts that (iii) nudges that cause people to attend to accuracy can increase veracity's role in sharing by increasing the probability that a_1=1 ($p(a_1=1)|\text{treatment} > p(a_1=1)|\text{control}$). That is, the accuracy nudge "shines an attentional spotlight" on the accuracy motive, increasing its chance to influence judgments.

3.3. Application of model to our setting

Next, we apply the general model presented in the previous section to the specific decision setting of our experiments. To do so, we consider k=3 content dimensions: to what extent the content seems inaccurate (*F*; 0=totally true to 1=totally false), aligned with the user's partisanship (*P*; from 0=totally misaligned to 1=totally aligned) or humorous (*H*, from 0=totally unfunny to 1=totally funny). There are, of course, numerous other relevant content dimensions that likely influence sharing which we do not include here; but in the name of tractability we focus on these dimensions as they are the dimensions that are manipulated in Studies 3 through 5 (article accuracy and partisanship are manipulated within-subjects in all experiments, accuracy focus is manipulated between-subjects in all experiments, and humor focus is manipulated between-subjects in Study 5). Below, we will demonstrate that modeling only these three dimensions allows us to characterize a large share of the variance in how often each headline gets shared; and look forward to future work building on the theoretical and experimental framework introduced here to explore a wider range of content dimensions.

We further stipulate that people are cognitively constrained to consider only m=2 of these dimensions in any given decision. We choose m=2 for the following reasons. First, the essence of the inattention-based account is that attention is limited, such that not all dimensions can be considered; thus, give that there are k=3 total dimensions, we necessarily choose a value of m<3 (m=3 gives the standard utility theory model, which by definition cannot account for the accuracy priming effects we demonstrate in our experiments). We choose m=2 over m=1 because it seems overly restrictive to assume that people can only consider a single dimension in any given situation. Furthermore, below we demonstrate that m=2 yields a better fit to our experimental data than m=1.

3.4. Analytic treatment of the effect of accuracy priming

In this section, we determine the impact of priming accuracy predicted by the preference-based versus inattention-based accounts. We define p as the probability that people do consider accuracy ($p(a_1=1)=p$). For simplicity, we assume that the two cases in which accuracy is considered are equally likely, such that people consider accuracy and partisanship ($a_1=a_2=1$ and

a_3=0) with probability p/2, and people consider accuracy and humor (a_1=a_3=1 and a_2=0) with probability p/2. Finally, with probability 1-p, people do not consider accuracy and instead consider partisanship and humor (a_1=0 and a_2=a_3=1). Also for simplicity, we use the simple decision rule whereby a piece of content x is shared if and only if $U(x) > 0$.

Within this setting, we can determine the probability that a given user (defined by her preferences β_F, β_P, and β_H, each of which is defined over the interval [-Inf,Inf]) shares a given piece of content x (defined by its characteristics $F(x)$, $C_2(x)$, and $C_3(x)$):

$$\underset{\substack{\text{(a) Considers accuracy \& partisanship}\\ a_1=1,\ a_2=1,\ a_3=0}}{\frac{p}{2}\,\mathbf{I}_{-\beta_F F(x)+\beta_P P(x)>0}} + \underset{\substack{\text{(b) Considers accuracy \& humor}\\ a_1=1,\ a_2=0,\ a_3=1}}{\frac{p}{2}\,\mathbf{I}_{-\beta_F F(x)+\beta_H H(x)>0}} + \underset{\substack{\text{(c) Considers partisanship \& humor}\\ a_1=0,\ a_2=1,\ a_3=1}}{(1-p)\,\mathbf{I}_{\beta_P P(x)+\beta_H H(x)>0}}$$

The key question, then, is how the users' sharing decisions vary with p, the probability that users' attentional spotlight is directed at accuracy. In particular, imagine a piece of content that is aligned with the users' partisanship $P(x)$=1 and humorous $H(x)$=1, but false $F(x)$=1. When the user does not consider accuracy (term c above, which occurs with probability 1-p), she will choose to share. When the user does consider accuracy (with probability p), her choice depends on her preferences. If β_F<β_P and β_F<β_H – that is, if the user cares about partisanship and humor more than accuracy, as per the preference-based account – she will still choose to share the misinformation. This is because the content's partisan alignment humorousness trumps its lack of accuracy, and therefore p does not impact sharing. Thus, if the sharing of misinformation is driven by a true lack of concern about veracity relative to other factors – as per the preference-based account – a manipulation that focuses attention on accuracy (and thereby increases p) will have no impact on the sharing of such misinformation.

If, on the other hand, β_F>β_P and/or β_F>β_H – that is, if the user cares about accuracy more than partisanship and/or humor – then directing attention at accuracy (and thereby increasing p) can influence sharing. If β_F>β_P, the user will choose not to share when considering accuracy and partisanship; and if β_F>β_H the user will choose not to share when considering accuracy and humor. As a result, increasing p will therefore decrease sharing. This scenario captures the essence of the inattention-based account.

Together, then, these two cases demonstrate how a manipulation that focuses attention on accuracy (increases p) – such as the manipulation in Studies 3 through 7 in the main text – will have differential impacts based on the relative importance the user places on accuracy. This illustrates how our experiments effectively disambiguate between the preference-based and inattention-based accounts of misinformation sharing.

This analysis also illustrates how drawing attention to a given dimension (e.g. priming it) need *not* translate into that dimension playing a bigger role in subsequent decisions. If the preference associated with that dimension is weak relative to the other dimensions (small β), then it will not drive choices even when attention is drawn to it. We will return to this observation when considering the lack of effect of the Active Control (priming humor) in Study 5.

3.5. Fitting the model to experimental data

In the previous section, we provided a conceptual demonstration of how the accuracy priming effect we observe empirically in Studies 3 through 7 is consistent with the inattention-based account and inconsistent with the preference-based account. Here, we take this further by fitting the model to experimental data. This allows us to directly test the predictions of the two accounts regarding various model parameters described above in Section 3.2, and thus to provide direct evidence for the role of inattention versus preferences in the sharing of misinformation. Fitting the model to the data also allows us to test how well our model can account for the observed patterns of sharing.

To perform the fitting, we use the pretest data for Studies 4 and 5 to calculate the average perceived accuracy ("What is the likelihood that the above headline is true", from 1 = Extremely unlikely to 7 = Extremely likely), political slant ("Assuming the above headline is entirely accurate, how favorable would it be to Democrats versus Republicans" from 1 = More favorable for Democrats to 5 = More favorable for Republicans), and humorousness ("In your opinion, is the above headline funny, amusing, or entertaining" from 1 = Extremely unfunny to 7 = Extremely funny) of each article. The pretest for the headlines in Studies 3 and 6 (both studies used the same items) did not include humorousness, and thus we cannot use those studies in the model fitting.

We must use the headline-level pretest ratings as a proxy for the ratings each individual would have of each article, because the participants in Studies 4 and 5 only made sharing decisions and did not rate each of the articles on perceived accuracy, political slant, or humorousness. Therefore, rather than separately estimating a model for every participant, we take a "representative agent" approach and estimate a single set of parameter values for the data averaged across subjects.

In order to define the political concordance $P(x)$ of headline x, however, it is necessary to consider Democrats and Republicans separately. This is because the extent to which a given headline is concordant for Democrats corresponds to the extent to which it is discordant for Republicans, and vice versa. Therefore, to create the dataset for fitting the model, the 44 total headlines (24 from Study 4 and 20 from Study 5) were each entered twice – once using the perceived accuracy ratings, humorousness ratings, and political slant ratings of Republican-leaning participants; and once using the perceived accuracy ratings, humorousness ratings, and 6 minus the political slant ratings (flipping the ratings to make them a measure of concordance) of Democratic-leaning participants. Each variable was scaled such that the minimum possible (rather than observed) value is 0 and the maximum possible (rather than observed) value is 1. This therefore yielded a set of 88 $\{F(x), P(x), H(x)\}$ value triples. For each of these 88 data points, we also calculated the corresponding average sharing intention in the control and in the treatment. (For maximum comparability across the two studies, we used the passive control not the active control, and the treatment not the importance treatment, in Study 5.)
We then determined the set of parameter values that minimized the mean-squared error (difference between the observed data and the model predictions), using a somewhat more complicated formulation that uses the more realistic logistic function for the decision rule

mapping from utility to choice, and allows each attentional case to have its own probability (rather than forcing the two cases that include accuracy to have the same probability):

$$p(share|control) = p_{1c}\frac{1}{1+\exp(-\theta(-\beta_F F(x)+\beta_P P(x)+k))} + p_{2c}\frac{1}{1+\exp(-\theta(-\beta_F F(x)+\beta_H H(x)+k))} + (1-p_{1c}-p_{2c})\frac{1}{1+\exp(-\theta(\beta_P P(x)+\beta_H H(x)+k))}$$

and

$$p(share|treatment) = p_{1t}\frac{1}{1+\exp(-\theta(-\beta_F F(x)+\beta_P P(x)+k))} + p_{2t}\frac{1}{1+\exp(-\theta(-\beta_F F(x)+\beta_H H(x)+k)} + (1-p_{1t}-p_{2t})\frac{1}{1+\exp(-\theta(\beta_P P(x)+\beta_H H(x)+k))}$$

Without loss of generality, as it is only the relative magnitude of the preference values that matters for choice, we fixed $\beta_F = 1$ and determined the best-fitting values of the remaining 8 parameters $\{\beta_P, \beta_H, p_{1c}, p_{2c}, p_{1t}, p_{2t}, \theta, k\}$, subject to the constraints $p_{1c}, p_{2c}, p_{1t}, p_{2t} \geq 0$, $p_{1c}, p_{2c}, p_{1t}, p_{2t} \leq 1$, $p_{1c} + p_{2c} \leq 1$, and $p_{1t} + p_{2t} \leq 1$. We did this by comparing the predicted probability of sharing from the model with the average sharing intention for each of the headline-level data points, and minimizing the MSE using the interior-point algorithm (as implemented by the function *fmincon* in Matlab R2018b). We performed this optimization beginning from 100 randomly selected initial parameter sets, and kept the solution with the lowest MSE.

We use the comparison of treatment and control data to disentangle preferences (β values) from attention (p-values). The key to our estimation strategy is that we hold the preference parameters β_F, β_P and β_H fixed across conditions while estimating different attention parameters in the control (p_{1c}, p_{2c}) and the treatment (p_{1t}, p_{2t}). As described above, fixed preferences is the standard assumption in virtually all utility theory models. Further evidence supporting the stability of the specifically relevant preference in our experiments comes from the observation, reported in the main text, that the treatment does not change participants' response to the post-experimental question about the importance of only sharing accurate content: If the treatment changed how much participants valued accuracy (rather than simply redirecting their attention), this would be likely to manifest itself as a greater reported valuation of accuracy.

We estimated the best-fit parameters separately for Studies 4 and 5. We did so for two reasons. First, the two studies were run with different populations (MTurk convenience sample versus Lucid quota-matched sample), so there is no reason to expect the best-fit parameter values to be the same. Second, because this analysis approach was not preregistered, we want to ensure that the results are replicable. Thus, we test the replicability of the results across the two studies, which differ in both the participants and the headlines used.

Finally, we estimate confidence intervals for the best-fit parameter values and associated quantities of interest, as well as p-values for relevant comparisons, using bootstrapping. Specifically, we construct bootstrap samples separately for each study by randomly resampling participants with replacement. For each bootstrap sample, we then use the rating-level data for the participants in the bootstrap sample to calculate mean sharing intentions for each headline in control and treatment, and then refit the model using these new sharing intentions values. We store the resulting best-fit parameters derived from 1500 bootstrap samples, and use the 2.5th percentile and 97.5th percentile of observed values to constitute the 95% confidence interval.

3.6. Results

We begin by examining the goodness of fit of our models, as the parameter estimates are only meaningful insomuch as the model does a good job of predicting the data. As mean-squared error (Study 4, MSE = 0.0036; Study 5, MSE = 0.0046) is not easily interpretable, we also consider the correlation between the model predictions and the observed average sharing intentions for each headline within each partisanship group (Democrats vs Republicans) in each experimental condition. As shown in Figure S1, we observe a high correlation in both Study 4, $r = 0.862$, and Study 5, $r = 0.797$. This indicates that despite only considering three of the many possible content dimensions, our model specification is able to capture much of the dynamics of sharing intentions observed in our experiments.

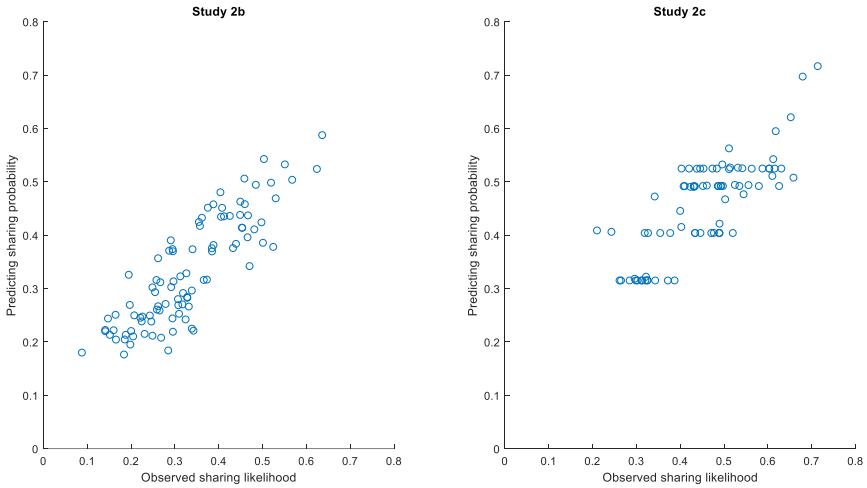


Figure S1. Observed and predicted sharing in Studies 4 and 5.

We now turn to the parameter estimates themselves. For each study, Extended Data Table 1 shows the best-fit parameter values; the overall probability that participants consider accuracy (p_{1c}+p_{2c}), political concordance (p_{1c}+(1-p_{1c}-p_{2c})), and humorousness (p_{2c}+(1-p_{1c}-p_{2c})) in the control; the overall probability that participants consider accuracy (p_{1t}+p_{2t}), political concordance (p_{1t}+(1-p_{1t}-p_{2t})), and humorousness (p_{2t}+(1-p_{1t}-p_{2t})) in the treatment; and the treatment effect on each of those quantities (probability in treatment minus probability in control). Note that because the best-fit values for β_H are substantially smaller than β_F (=1) and β_P – that is, because participants don't put much value on humorousness – the estimates for probability of considering humorousness are not particularly meaningful. This is because even if participants did pay attention to humorousness, it would always be outweighed by whichever other factor was being

considered; and thus it is not possible from the choice data to precisely determine whether humorousness was attended to; this is not problematic for us, however, as none of the key predictions involve probability of attending to humorousness.

There are three key results in Extended Data Table 1. First, inconsistent with the preference-based account, the best-fit preference parameters indicate that participants value accuracy as much as or more than partisanship. Thus, they would be unlikely to share false but politically concordant content if they were attending to accuracy and partisanship. (This is not to say that partisanship is unimportant, but rather that partisanship does not *override* accuracy – ideologically aligned content must *also* be sufficiently accurate in order to have a high sharing probability). Second, the best-fit attention parameters indicate participants often fail to consider accuracy because their attention is directed to other content dimensions. This can lead them to share content that they would have assessed as inaccurate (and chosen not to share), had they considered accuracy. And finally, the Treatment increases participants' likelihood of considering accuracy (and thereby reduces the sharing of false statements).

3.7. Alternative model specifications

In this section, we compare the performance of our model to various alternative specifications. First, we contrast our assumption that participants can attend to m=2 of the k=3 content dimensions in any given decision with a model in which m=1 (i.e. where participants can only consider one dimension per decision). This yields the following formulation:

$$p(share|control) = p_{1c}\frac{1}{1+\exp(-\theta(-\beta_F F(x)+k))} + p_{2c}\frac{1}{1+\exp(-\theta(\beta_P P(x)+k))} + (1-p_{1c} - p_{2c)}\frac{1}{1+\exp(-\theta(\beta_H H(x)+k))}$$

and

$$p(share|treatment))) = p_{1t}\frac{1}{1+\exp(-\theta(-\beta_F F(x)+k))} + p_{2t}\frac{1}{1+\exp(-\theta(\beta_P P(x)+k))} + (1-p_{1t} - p_{2t)}\frac{1}{1+\exp(-\theta(\beta_H H(x)+k))}$$

Since this formulation has the same number of free parameters as the main m=2 model, it is straightforward to compare model fit by simply asking which model fits the data better. Fitting this model to the data yields a higher mean-squared error than our main model with m=2 in both Study 4 (MSE=0.0043 vs MSE=0.0036 in the m=2 model) and Study 5 (MSE=0.0056 vs MSE=0.0046 in the m=2 model), indicating that the m=2 model is preferable.

Next, we contrast our model – based on cognitive constraints – with the formulation used in prior models of attention and preferences [9,10] in which all preferences are considered in every decision, but are differentially weighted by attention. This alternative approach yields the following formulation:

$$p(share|control) = \frac{1}{1 + \exp(-\theta(-p_{1c}\beta_F F(x) + p_{2c}\beta_P P(x) + (1 - p_{1c} - p_{2c})\beta_H H(x) + k))}$$

and

$$p(share|treatment) = \frac{1}{1 + \exp(-\theta(-p_{1t}\beta_F F(x) + p_{2t}\beta_P P(x) + (1 - p_{1t} - p_{2t})\beta_H H(x) + k))}$$

Once again, this alternative formulation has the same number of free parameters as our main model, allowing for straightforward model comparison. Fitting this model to the data yields a higher mean-squared error than our main model in both Study 4 (MSE=0.0039 vs MSE=0.0036 in the main model) and Study 5 (MSE=0.0057 vs MSE=0.0046 in the main model), indicating that the main model is preferable.

Next, we examine the simplifying assumption in our main model that attention (i.e. the probability that any given content dimension is considered) is exogenously determine (e.g. by the context). In reality, one's preferences may also influence how one allocates one's attention. For example, a person who cares a great deal about accuracy may be more likely to attend to accuracy. To consider the consequences of such a dependence, we additionally weight each attention scenario not just be its associated value of p (p_{1c}, p_{2c}, etc.) but also by the relative preference weight put on the two dimensions considered in that scenario. This yields the following formulation:

$$\begin{aligned} p(share|control) \\ &= p_{1c}\frac{\beta_F + \beta_P}{\pi_c}\frac{1}{1 + \exp(-\theta(-\beta_F F(x) + k))} \\ &+ p_{2c}\frac{\beta_F + \beta_H}{\pi_c}\frac{1}{1 + \exp(-\theta(\beta_P P(x) + k))} + (1 - p_{1c} \\ &- p_{2c})\frac{\beta_P + \beta_H}{\pi_c}\frac{1}{1 + \exp(-\theta(\beta_H H(x) + k))} \end{aligned}$$

and

$$\begin{aligned} p(share|treatment) \\ &= p_{1t}\frac{\beta_F + \beta_P}{\pi_t}\frac{1}{1 + \exp\big(-\theta(-\beta_F F(x) + k)\big)} \\ &+ p_{2t}\frac{\beta_F + \beta_H}{\pi_t}\frac{1}{1 + \exp\big(-\theta(\beta_P P(x) + k)\big)} + (1 - p_{1t} \\ &- p_{2t})\frac{\beta_P + \beta_H}{\pi_t}\frac{1}{1 + \exp(-\theta(\beta_H H(x) + k))} \end{aligned}$$

where π_c and π_t are normalization constants that force the probabilities to sum to one, such that

$$\pi_c = p_{1c}(\beta_F + \beta_P) + p_{2c}(\beta_F + \beta_H) + (1 - p_{1c} - p_{2c})(\beta_P + \beta_H)$$
$$\pi_t = p_{1t}(\beta_F + \beta_P) + p_{2c}(\beta_F + \beta_H) + (1 - p_{1c} - p_{2c})(\beta_P + \beta_H)$$

Once again, this model has the same number of free parameters as the main model. Unlike the previous alternative models, this model of endogenous attention fits the data exactly as well as

the main (exogenous attention) model (identical MSE to 4 decimal places in both studies). The resulting fits have identical preference values of β_P and β_H (and therefore contradict the preference-based account in the same way as the main model) and qualitatively similar results regarding the attention parameters: in the control, participants often fail to consider accuracy (overall probability of considering accuracy = 0.07 in Study 4, 0.61 in Study 5), and the treatment increases participants' probability of considering accuracy (by 0.02 in Study 4, and 0.08 in Study 5). Furthermore, an analytic treatment of this model provides equivalent results to the analysis of the exogenous attention model presented above in Section 3.4.

4. Ethics of Digital Field Experimentation

Field experimentation, such as our Study 7, necessarily involves engaging in people's natural activities to assess the effect of a treatment *in situ*. As digital experimentation on social media becomes more attractive to social scientists, there are increasing ethical considerations that must be taken into account [19–21].

One such consideration is the nature of the interaction between Twitter users and our bot accounts. As discussed above, this involved following individuals who shared links to misinformation sites, and then sending a DM to those individuals who followed our bot accounts back. We believe that the potential harm of an account following and sending a DM to an individual is minimal; and that the potential benefits of scientific understanding and an increase in shared news quality outweigh that negligible risk. Both the Yale University Committee for the Use of Human Subjects (IRB protocol #2000022539) and the MIT COUHES (Protocol #1806393160) agreed with our assessment. With regard to informed consent, it is standard practice in field experiments to eschew informed consent because much of the value of field experiments comes from participants not knowing they are in an experiment (thus providing ecological validity). As obtaining informed consent would disrupt the user's normal experience using Twitter, and greatly reduce the validity of the design – and the risks were minimal – both institutional review boards waived the need for informed consent. A final consideration is the ethical collection of individuals' tweet histories for analysis. Since we are only considering publicly available tweets, and hence any collated dataset would be the product of secondary research, we believe this to be an acceptable practice.

There is the open question of how these considerations interact, and if practices that are separately appropriate can create ethically ambiguous situations when conducted conjointly. Data rights on social media are a complicated and ever-changing social issue with no clear answers. We hope Study 7 highlights some principles and frameworks for considering these issues in the context of digital experimentation, and helps create more discussion and future work on concretely establishing norms of engagement.

There has been some discussion about the ethics of nudges, primes, modifications to choice architectures, and other interventions for digital behavior change. Some worry that these interventions can be paternalistic, and favor the priorities of platform designers over users. Our intervention - making the concept of accuracy salient - does not prescribe any agenda or normative stance to users. We do not tell users what is accurate versus inaccurate, or even tell them that they should be taking accuracy into account when sharing. Rather, the intervention simply moves the spotlight of attention towards accuracy, and then allows the user to make their own determination of accuracy and make their own choice about how to act on that determination.

While we believe this intervention is ethically sound, we also acknowledge the fact that if this methodology was universalized as a new standard for social science research, it could further dilute and destabilize the Twitter ecosystem, which already suffers from fake accounts, spam, and misinformation. Future work should invest in new frameworks for digital experimentation that maintains social media's standing as a town square for communities to genuinely engage in communication, while also allowing researchers to causally understand user behavior on the platform. These frameworks may involve, for example, external software libraries built on top of publicly available APIs, or explicit partnerships with the social media companies themselves.

5. Additional Analysis for Study 7

Table S6 shows a consistent significant interaction between treatment and the tweet being an RT-without-comment, such that the treatment consistently increases the average quality of RTs-without-comment but has no significant effect on primary tweets. (We do not conduct this interaction analysis for summed relative quality or discernment, because the differences in tweet volume between RTs-without-comment and primary tweets makes those measures not comparable.)

Randomization-Failure	Article Type	Model Spec	Interaction			Simple effect on NRT			Simple effect on RT		
			Coeff	Reg p	FRI p	Coeff	Reg p	FRI p	Coeff	Reg p	FRI p
ITT	All	Wave FE	0.008	**0.004**	**0.003**	0.000	0.741	0.701	0.007	**0.003**	**0.004**
ITT	All	Wave PS	0.008	**0.006**	**0.003**	0.000	0.761	0.756	0.007	**0.004**	**0.004**
ITT	All	Date FE	0.007	**0.022**	**0.004**	-0.001	0.725	0.800	0.006	**0.017**	**0.014**
ITT	All	Date PS	0.007	**0.031**	**0.006**	-0.001	0.725	0.703	0.006	**0.027**	**0.012**
Exclude	All	Wave FE	0.008	**0.006**	**0.004**	-0.001	0.676	0.635	0.007	**0.004**	**0.009**
Exclude	All	Wave PS	0.008	**0.007**	**0.004**	-0.001	0.690	0.687	0.007	**0.005**	**0.008**
Exclude	All	Date FE	0.006	**0.033**	**0.007**	-0.001	0.629	0.740	0.006	**0.032**	**0.032**
Exclude	All	Date PS	0.006	**0.040**	**0.009**	-0.001	0.629	0.653	0.006	**0.042**	**0.023**
ITT	No Opinion	Wave FE	0.009	**0.001**	**0.001**	-0.001	0.466	0.453	0.008	**0.001**	**0.003**
ITT	No Opinion	Wave PS	0.009	**0.001**	**0.001**	-0.001	0.454	0.491	0.008	**0.002**	**0.004**
ITT	No Opinion	Date FE	0.008	**0.007**	**0.001**	-0.001	0.458	0.646	0.007	**0.009**	**0.013**
ITT	No Opinion	Date PS	0.008	**0.009**	**0.001**	-0.001	0.458	0.493	0.007	**0.013**	**0.007**
Exclude	No Opinion	Wave FE	0.009	**0.001**	**0.002**	-0.001	0.476	0.486	0.008	**0.001**	**0.009**
Exclude	No Opinion	Wave PS	0.009	**0.002**	**0.001**	-0.001	0.450	0.505	0.008	**0.003**	**0.008**
Exclude	No Opinion	Date FE	0.007	**0.013**	**0.002**	-0.001	0.442	0.655	0.006	**0.017**	**0.029**
Exclude	No Opinion	Date PS	0.008	**0.015**	**0.001**	-0.001	0.442	0.495	0.006	**0.021**	**0.014**

Table S6. Coefficients and p-values associated with the interaction between treatment and tweet type, and each simple effect of treatment, when predicting average relative quality for Study 7. In the model specification column, FE represents fixed effects (i.e. just dummies) and PS represents post-stratification (i.e. centered dummies interacted with the post-treatment dummy). P-values below 0.05 are bolded.

The analyses presented in Extended Data Table 4 collapse across waves to maximize statistical power. As evidence that this aggregation is justified, we examine the models in which the treatment effect is post-stratified on wave (i.e. the wave dummies are interacted with the post-treatment dummy). Table S7 shows the p-values generated by a joint significance test over the wave-post-treatment interactions (i.e. testing whether the treatment effect differed significantly in size across waves) for the four dependent variables crossed with the four possible inclusion criteria choices. As can be seen, in all cases the joint significance test is extremely far from significant. This lack of significant interaction between treatment and wave supports our decision to aggregate the data across waves.

Tweet Type	Randomization-Failure	Average Relative Quality	Summed Relative Quality	Discernment
All	Exclude	0.685	0.378	0.559
All	ITT	0.743	0.313	0.613
RT	Exclude	0.710	0.508	0.578
RT	ITT	0.722	0.535	0.687

Table S7. P-values generated by a joint significant test of the interaction between wave2 and post-treatment and wave3 and post-treatment, from the models in Extended Data Table 4 where treatment effect is post-stratified on wave.

Next, Table S8 shows models testing for an interaction between the treatment and the user's number of followers (log-transformed due to extreme right skew) when predicting average relative quality of tweets. As can be seen, none of the interactions are significant, and the sign of all interactions is positive. Thus, there is no evidence that the treatment is less effective for users with more followers. If anything, the effect is directionally in the opposite direction.

Tweet Type	Article Type	Randomization-Failure	Model Spec	Coeff	Reg p	FRI p
All	All	ITT	Wave FE	0.003	0.252	0.905
All	All	ITT	Wave PS	0.003	0.200	0.123
All	All	ITT	Date FE	0.002	0.360	0.301
All	All	ITT	Date PS	0.002	0.468	0.441
RT	All	ITT	Wave FE	0.002	0.364	0.919
RT	All	ITT	Wave PS	0.002	0.319	0.201
RT	All	ITT	Date FE	0.002	0.468	0.375
RT	All	ITT	Date PS	0.002	0.452	0.455
All	All	Exclude	Wave FE	0.004	0.143	0.977
All	All	Exclude	Wave PS	0.004	0.124	0.066
All	All	Exclude	Date FE	0.003	0.225	0.152
All	All	Exclude	Date PS	0.003	0.357	0.324
RT	All	Exclude	Wave FE	0.003	0.215	0.979
RT	All	Exclude	Wave PS	0.003	0.204	0.111
RT	All	Exclude	Date FE	0.003	0.307	0.200
RT	All	Exclude	Date PS	0.003	0.345	0.354
All	No Opinion	ITT	Wave FE	0.003	0.190	0.954
All	No Opinion	ITT	Wave PS	0.004	0.167	0.121
All	No Opinion	ITT	Date FE	0.003	0.285	0.216
All	No Opinion	ITT	Date PS	0.002	0.380	0.386
RT	No Opinion	ITT	Wave FE	0.002	0.296	0.956
RT	No Opinion	ITT	Wave PS	0.003	0.269	0.202
RT	No Opinion	ITT	Date FE	0.002	0.403	0.334
RT	No Opinion	ITT	Date PS	0.002	0.371	0.458
All	No Opinion	Exclude	Wave FE	0.004	0.098	0.986
All	No Opinion	Exclude	Wave PS	0.004	0.097	0.064
All	No Opinion	Exclude	Date FE	0.004	0.161	0.094
All	No Opinion	Exclude	Date PS	0.003	0.275	0.286
RT	No Opinion	Exclude	Wave FE	0.003	0.179	0.984
RT	No Opinion	Exclude	Wave PS	0.003	0.178	0.128
RT	No Opinion	Exclude	Date FE	0.003	0.264	0.173
RT	No Opinion	Exclude	Date PS	0.003	0.283	0.375

Table S8. Coefficients and p-values associated with the interaction between treatment and log(# followers) each model predicting average relative quality for Study 3. In the model specification column, FE represents fixed effects (i.e. just dummies) and PS represents post-stratification (i.e. centered dummies interacted with the post-treatment dummy). All p-values are above 0.05.

Finally, as shown in Table S9, we see no evidence of a treatment effect when considering tweets that did not contain links to any of the rated news sites, or when considering the probability that any rated tweets occurred.

Tweet Type	Randomization-Failure	Model Spec	Tweets without rated links			Any rated tweets		
			Coeff	Reg p	FRI p	Coeff	Reg p	FRI p
All	ITT	Wave FE	0.492	0.483	0.342	0.004	0.600	0.602
All	ITT	Wave PS	0.364	0.577	0.460	0.004	0.611	0.590
All	ITT	Date FE	0.160	0.843	0.788	0.006	0.472	0.494
All	ITT	Date PS	0.126	0.873	0.825	0.010	0.293	0.181
All	Exclude	Wave FE	0.221	0.756	0.672	-0.001	0.890	0.894
All	Exclude	Wave PS	0.150	0.823	0.763	0.000	0.978	0.979
All	Exclude	Date FE	-0.232	0.779	0.697	0.001	0.929	0.929
All	Exclude	Date PS	-0.127	0.876	0.827	0.006	0.500	0.397
RT	ITT	Wave FE	0.440	0.408	0.266	-0.001	0.917	0.895
RT	ITT	Wave PS	0.332	0.495	0.367	-0.002	0.806	0.760
RT	ITT	Date FE	0.246	0.687	0.569	0.001	0.943	0.927
RT	ITT	Date PS	0.139	0.814	0.744	0.004	0.657	0.560
RT	Exclude	Wave FE	0.266	0.620	0.505	-0.006	0.455	0.338
RT	Exclude	Wave PS	0.197	0.692	0.600	-0.006	0.450	0.346
RT	Exclude	Date FE	-0.004	0.995	0.992	-0.005	0.599	0.510
RT	Exclude	Date PS	-0.028	0.963	0.946	0.001	0.928	0.907

Table S9. Coefficients and p-values associated with each model predicting number of unrated tweets and presence of any rated tweets for Study 7. In the model specification column, FE represents fixed effects (i.e. just dummies) and PS represents post-stratification (i.e. centered dummies interacted with the post-treatment dummy).

Turning to visualization, in Figure S2 we show the results of domain-level analyses. These analyses compute the fraction of pre-treatment rated links that link to each of the 60 rated domains, and the fraction of rated links in the 24 hours post-treatment that link to each of the 60 rated domains. For each domain, we then plot the difference between these two fractions on the y-axis, and the fact-checker trust rating from Pennycook & Rand[22] on the x-axis.

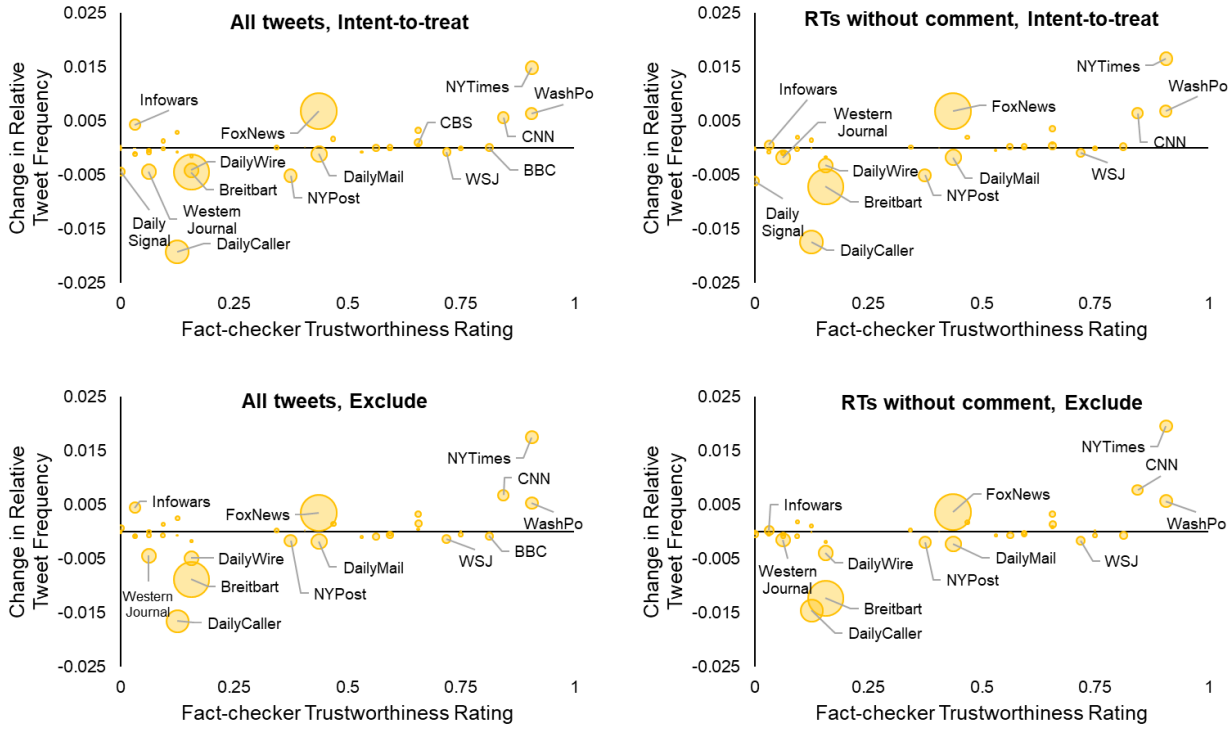


Figure S2. Domain-level analysis for each combination of approach to randomization failure (exclusion or intent-to-treat) and tweet type (all or only RTs-without-comment). Size of dots is proportional to pre-treatment tweet count. Outlets with at least 500 pre-treatment tweets are labeled.

6. Modeling the spread of misinformation

Our paper theoretically and empirically investigates the role of accuracy and inattention in individuals' decisions about what to share online. To investigate how these individual-level choices – and the accuracy nudge interventions we introduce to improve such choices – translate into population-level outcomes regarding the spread of misinformation, we employ simulations of social spreading dynamics. The key goal of the simulations is to shed light on how network effects either suppress or amplify the impact of the accuracy intervention (which we have shown to improve individual choices).

In our simulations, a population of agents is embedded in a network. When an agent is first exposed to a piece of information, they share it with probability s. Based on the Control conditions of Study 3-6, we take the probability of sharing a piece of fake news at baseline (i.e., without intervention) to be approximately s=0.3. The Full Attention Treatment of Study 6 indicates that if an intervention was able to entirely eliminate inattention, the probability of sharing would be reduced by 50%. Thus, we vary s across the interval [0.15, 0.3] and examine the impact on the spread of misinformation. If an agent does choose to share a piece of information, each of their followers is exposed to that information with probability p (p<<1, as most shared content is never seen because it is quickly pushed down the newsfeed queue [23]; we use p=0.1).

In each run of the simulation, a piece of misinformation is seeded in the network by randomly selecting an initial user to be exposed to that misinformation. They then decide whether to share based on s, if they do then each of their followers is exposed with probability p; then each of the exposed followers shares with probability s, and if so then their followers are exposed with probability p, and so on. The simulation then runs until no new exposures occur, and the total fraction of the population exposed to the piece of information across the simulation run is calculated.

This procedure thus allows us to determine how a given decrease in individuals' probability of sharing misinformation impacts the population-level outcome of misinformation spread. We examine how the fraction of agents that get exposed varies with the magnitude of the intervention effect (extent to which s is reduced), the type of network structure (cycle, Watts-Strogatz small-world network with rewiring rate of 0.1, or Barabási–Albert scale-free network), and the density of the network (average number of neighbors k). Our simulations use a population of size N=1000, and we show the average result of 10,000 simulation runs for each set of parameter values.

Extended Data Figure 6 shows how a given percentage reduction in individual sharing probability (between 0 and 50%) translates into a percentage reduction in the fraction of the population that is exposed to the piece of misinformation.

7. Supplementary references

1. Fishburn, P. C. Utility Theory. *Manage. Sci.* **14**, 335–378 (1968).
2. Stigler, G. J. The Development of Utility Theory. I. *J. Polit. Econ.* **58**, 307–327 (1950).
3. Quiggin, J. A theory of anticipated utility. *J. Econ. Behav. Organ.* **3**, 323–343 (1982).
4. Barberis, N. C. Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives* **27**, 173–196 (2013).
5. Taylor, S. E. & Thompson, S. C. Stalking the elusive 'vividness' effect. *Psychol. Rev.* **89**, 155–181 (1982).
6. Ajzen, I. Nature and Operation of Attitudes. *Annu. Rev. Psychol.* **52**, 27–58 (2001).
7. Simon, H. A. & Newell, A. Human problem solving: The state of the theory in 1970. *Am. Psychol.* **26**, 145–159 (1971).
8. Higgins, E. T. Knowledge activation: Accessibility, applicability, and salience. in *Social Psychology: Handbook of Basic Principles* (eds. Higgins, E. T. & Kruglanski, A. W.) 133–168 (Guilford Press, 1996).
9. Bordalo, P., Gennaioli, N. & Shleifer, A. Salience Theory of Choice Under Risk. *Q. J. Econ.* **127**, 1243–1285 (2012).
10. Koszegi, B. & Szeidl, A. A Model of Focusing in Economic Choice. *Q. J. Econ.* **128**, 53–104 (2012).
11. Camerer, C. F., Loewenstein, G. & Rabin, M. *Advances in Behavioral Economics*. (Princeton University Press, 2004).
12. Evans, J. S. B. T. & Stanovich, K. E. Dual-process theories of higher cognition: Advancing the debate. *Perspect. Psychol. Sci.* **8**, 223–241 (2013).
13. Fiske, S. & Taylor, S. *Social cognition: From brains to culture*. (McGraw-Hill, 2013).
14. Simon, H. Theories of bounded rationality. in *Decision and Organization* 161–176 (1972).
15. Stahl, D. O. & Wilson, P. W. On players′ models of other players: Theory and experimental evidence. *Games Econ. Behav.* **10**, 218–254 (1995).
16. Stanovich, K. E. *The robot's rebellion: Finding meaning in the age of Darwin*. (Chicago University Press, 2005).
17. Pennycook, G., Fugelsang, J. A. & Koehler, D. J. What makes us think? A three-stage dual-process model of analytic engagement. *Cogn. Psychol.* **80**, 34–72 (2015).
18. Lewandowsky, S., Ecker, U. K. H. & Cook, J. Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era. *J. Appl. Res. Mem. Cogn.* **6**, 353–369 (2017).
19. Gallego, J., Martínez, J. D., Munger, K. & Vásquez-Cortés, M. Tweeting for peace: Experimental evidence from the 2016 Colombian Plebiscite. *Elect. Stud.* **62**, 102072 (2019).

20. Desposato, S. *Ethics and experiments: Problems and solutions for social scientists and policy professionals*. (Routledge, 2015).

21. Taylor, S. J. & Eckles, D. Randomized experiments to detect and estimate social influence in networks. in *Complex Spreading Phenomena in Social Systems* (eds. Lehmann, S. & Ahn, Y. Y.) 289–322 (Springer, 2018).

22. Pennycook, G. & Rand, D. G. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc. Natl. Acad. Sci.* (2019). doi:10.1073/pnas.1806781116

23. Hodas, N. O. & Lerman, K. The simple rules of social contagion. *Sci. Rep.* **4**, 1–7 (2014).

nature research

Corresponding author(s): Gordon Pennycook, David Rand

Last updated by author(s): Jan 16, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data collection was completed using Qualtrics survey software. Data collection and intervention implementation for Study 7 was done in Python 3.6. Matlab version R2018b was used for fitting the limited attention utility model to the data.
Data analysis	Data analysis was done in R 4.0.2 and Stata 16.1.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data and materials for Studies 1 through 6 are available at https://osf.io/p6u8k/. Due to privacy concerns, data from Study 7 are available upon request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Study designs are outlined in the Methods section. Studies 1-6 involve presenting people with true and false news headlines (in a Facebook format) and asking them if they would consider sharing them on social media (and, in some cases, whether they believe the headlines to be accurate). Study design for Study 7 involved messaging Twitter users with an accuracy prompt and tracking their subsequent Twitter behavior.
Research sample	Studies 1-6 involved participants from Amazon's Mechanical Turk and Lucid. The former is not nationally representative; the latter is quote-matched to be representative based on age, gender, ethnicity, and region. Study 7 involved Twitter users and was not representative. Our survey samples are "convenience samples". However, given that our focus is on social media behavior, recruiting participants who complete online studies is a good fit. Demographics for the samples can be found in our Methods.
Sampling strategy	Sample sizes for Studies 1-6 were based on previous studies of this nature and the maximum amount of money that we wished to spend on the studies, given that we had no basis for an a priori estimated effect size (and such power analyses are often arbitrary anyway: http://datacolada.org/4). Study 37involved following users on Twitter, and to comply with Twitter, our subject pool was the individuals who followed back our accounts.
Data collection	Data for Studies 1-6 was collected using online survey software (Qualtrics), which completed the randomization into separate experimental conditions. Study 7 data was taken from Twitter.
Timing	Study 1: August 13, 2019. Study 2: January 9-13, 2020. Study 3: October 4-6, 2017. Study 4: November 28-30, 2017. Study 5: May 1, 2019. Study 6: August 24, 2017. Study 7: 4/20/2018-4/27/2018 (Wave 1), 9/12/2018-9/14/2018 (Wave 2) and 1/28/2019-2/08/2019 (Wave 3).
Data exclusions	Participants in Studies 3-6 were excluded if they did not use social media or had no interest in sharing political content. Study 3 involved a variety of exclusions that are explained in detail in the supplementary document (e.g., those who didn't tweet links to the websites on our fact-checking rating list were excluded as they did not provide any data). This was preregistered. See Methods for inclusion criteria for Study 7.
Non-participation	Dropout was low in Studies 1-6. Study 1: 1.5% of the full sample. Study 2: 0%. Study 3: 5.7%. Study 4: 5.5%. Study 5: 7.5%. Study 6: 2.4%. Dropout was not possible in Study 7.
Randomization	For Studies 1-6, randomization was completed using Qualtrics survey software. For Study 7, a randomization schedule was computed to assign participants to experimental conditions while accounting for the blocking procedure. Then the Direct Messages were sent out through the Twitter API using this randomization schedule.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☐	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above.
Recruitment	Participants in Studies 1-6 were recruited online via Amazon's Mechanical Turk and Lucid for Academics. Study 7 participants were selected via previous engagement with low-quality news sources. Regarding self-selection bias: For our survey experiments, participants have to decide to complete the survey. This is, of course, true for any survey study run that involves participant consent. In the Twitter experiment, only those who "followed back" our bots were included in the study; nonetheless, we randomized our treatment within this group and therefore our causal inference is maintained.
Ethics oversight	We had research clearance from Yale Human Subjects Committee (Studies 3, 4, 6, 7), MIT Committee on the Use of Humans as Experimental Subjects (Study 2, 7), and the University of Regina Research Ethics board (Studies 1, 5). Participants in Studies 3, 4, and 6 were not given consent forms because the Yale IRB had a consent form exception for online studies. Consent was not obtained for Study 7, consistent with our approved ethics protocols (see supplementary materials, Section 4 for further explanation).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☒	☐	Public health
☒	☐	National security
☒	☐	Crops and/or livestock
☒	☐	Ecosystems
☒	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☒	☐	Demonstrate how to render a vaccine ineffective
☒	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	☐	Increase transmissibility of a pathogen
☒	☐	Alter the host range of a pathogen
☒	☐	Enable evasion of diagnostic/detection modalities
☒	☐	Enable the weaponization of a biological agent or toxin
☒	☐	Any other potentially harmful combination of experiments and agents