AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2023, Vol. 152, No. 8, 2205–2236
https://doi.org/10.1037/xge0001381

Truth Sensitivity and Partisan Bias in Responses to Misinformation

Bertram Gawronski, Nyx L. Ng, and Dillon M. Luke
Department of Psychology, University of Texas at Austin

Misinformation represents one of the greatest challenges for the functioning of societies in the information age. Drawing on a signal-detection framework, the current research investigated two distinct aspects of misinformation susceptibility: *truth sensitivity*, conceptualized as accurate discrimination between true and false information, and *partisan bias*, conceptualized as lower acceptance threshold for ideology-congruent information compared to ideology-incongruent information. Four preregistered experiments ($n = 2{,}423$) examined (a) truth sensitivity and partisan bias in veracity judgments and decisions to share information and (b) determinants and correlates of truth sensitivity and partisan bias in responses to misinformation. Although participants were able to distinguish between true and false information to a considerable extent, sharing decisions were largely unaffected by actual information veracity. A strong partisan bias emerged for both veracity judgments and sharing decisions, with partisan bias being unrelated to the overall degree of truth sensitivity. While truth sensitivity increased as a function of cognitive reflection during encoding, partisan bias increased as a function of subjective confidence. Truth sensitivity and partisan bias were both associated with misinformation susceptibility, but partisan bias was a stronger and more reliable predictor of misinformation susceptibility than truth sensitivity. Implications and open questions for future research are discussed.

Public Significance Statement
This research investigated two distinct aspects of misinformation susceptibility: (a) the ability to accurately distinguish between true and false information (*truth sensitivity*) and (b) the tendency to accept information that is congruent with one's ideological beliefs and dismiss information that is incongruent with one's ideological beliefs (*partisan bias*). The findings suggest that interventions to reduce misinformation susceptibility should adopt a multi-faceted approach targeting both factors.

Keywords: misinformation, partisan bias, signal detection, social media, truth discernment

Supplemental materials: https://doi.org/10.1037/xge0001381.supp

This article was published Online First March 27, 2023.

Bertram Gawronski https://orcid.org/0000-0001-7938-3339

We thank Ana Sofia Cerda Ibarrola, Timothy Jackson, Audrey-Anna Scalici, and Alyssa Yarbrough for their help with the preparation of the study materials, and Alyssa Yarbrough for her assistance with the data collection. This research was supported by National Science Foundation Grant BCS-2040684. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors declare no conflict of interest.

The data, analysis codes, research materials, and links to preregistrations of the reported studies are available at https://osf.io/d2rne/.

Bertram Gawronski served as lead for conceptualization, data curation, formal analysis, funding acquisition, supervision, and writing–original draft. Nyx L. Ng served as lead for investigation and served in a supporting role for data curation, formal analysis, and supervision. Dillon M. Luke served in a supporting role for data curation and investigation. Nyx L. Ng and Dillon M. Luke contributed to writing–review and editing equally.

Correspondence concerning this article should be addressed to Bertram Gawronski, Department of Psychology, University of Texas at Austin, 108 E Dean Keeton A8000, Austin, TX 78712, United States. Email: gawronski@utexas.edu

One of the greatest challenges for the functioning of societies in the information age is the prevalence and spread of misinformation (Lewandowsky et al., 2012). As people base their decisions on information that is available to them, inaccurate information can lead to suboptimal decision outcomes even when the decision-process itself is perfectly rational from a normative point of view (Trafimow, 2015). In some cases, suboptimal outcomes of misinformed decisions affect not only the individual who made the decision, but society as a whole. For example, during the COVID-19 pandemic, misinformation has led some people to engage in behaviors that put others' lives at risk by contributing to the spread of the coronavirus. In the political domain, concerns have been raised that misinformation poses a major threat to democratic societies because of its potential to undermine people's trust in democratic institutions.

A central question in psychological research on misinformation is how people determine whether a given piece of information is true or false (Brashier & Marsh, 2020). In the current research, we used signal detection theory (SDT; Green & Swets, 1966) to investigate two distinct aspects of responses to political misinformation: (a) *truth sensitivity*, conceptualized as accurate discrimination between true and false information, and (b) *partisan bias*, conceptualized as lower acceptance threshold for ideology-congruent information

compared to ideology-incongruent information (see Batailler et al., 2022).[1] In a series of four preregistered experiments, we examined (a) truth sensitivity and partisan bias in veracity judgments and decisions to share information online, and (b) determinants and correlates of truth sensitivity and partisan bias in responses to misinformation. In an integrative analysis of the combined data from the four studies, we also explored the extent to which susceptibility to misinformation is accounted for by truth sensitivity and partisan bias, respectively.

Why Do People Fall for Misinformation?

Although susceptibility to misinformation has been linked to a broad range of psychological factors (Lewandowsky et al., 2012), current debates about why people fall for misinformation are dominated by two competing theoretical views. Motivational accounts suggest that susceptibility to misinformation is primarily caused by processes of motivated reasoning (see Kruglanski et al., 2020; Kunda, 1990), producing a pattern of partisan bias in responses to true and false information. According to social-identity accounts (e.g., Van Bavel & Pereira, 2018), people are motivated to support and protect beliefs that are central to their identity, leading them to accept ideology-congruent information and reject ideology-incongruent information irrespective of the actual veracity of the information. For example, supporters of a particular politician may accept false information that reflects positively on that politician and dismiss true information if it reflects negatively on that politician. Conversely, critics of the same politician may accept false information that reflects negatively on that politician and dismiss true information if it reflects positively on that politician.

Different from motivational accounts, cognitive accounts suggest that susceptibility to misinformation is caused by insufficient cognitive reflection rather than partisan bias (e.g., Pennycook, 2023; Pennycook & Rand, 2021). According to cognitive accounts, failures to distinguish between true and false information are the product of shallow processing, which can lead people to mistakenly accept false information as true and mistakenly reject true information as false. Importantly, the proposed effect of cognitive reflection is assumed to be independent of partisanship, in that shallow processing undermines truth discernment for both ideology-congruent and ideology-incongruent information. Applied to the above example, these ideas suggest that a person's accuracy in distinguishing between true and false information about a particular politician may be low due to shallow cognitive processing regardless of whether the person supports or opposes that politician.

Identification of Misinformation as a Signal Detection Problem

Although cognitive and motivational accounts have been discussed as competing explanations for misinformation susceptibility (e.g., Pennycook & Rand, 2021), an analysis in terms of SDT suggests that truth sensitivity and partisan bias are not mutually exclusive, but instead reflect two distinct factors in responses to misinformation (Batailler et al., 2022). To illustrate how truth sensitivity and partisan bias are reflected in responses to true and false information, we first describe SDT's concepts of *discrimination sensitivity* and *response threshold* and then explain how truth sensitivity and partisan bias can be understood in terms of SDT.

SDT is a mathematical approach to understanding how different factors influence people's ability to distinguish signal from noise (Green & Swets, 1966). This approach is content-independent in the sense that it can be applied to any domain where people make binary judgments about two classes of stimuli. Responses in such binary classification problems can be described with a 2×2 matrix capturing four potential outcomes (see Table 1). For example, applied to the question of how people distinguish between true and false information, correct classification of true information as true can be described as *hit*; correct classification of false information as false can be described as *correct rejection*; incorrect classification of true information as false can be described as *miss*; and incorrect classification of false information as true can be described as *false alarm*. Drawing on this conceptualization, research on why people fall for misinformation can be understood as being concerned with the causes of false alarms, that is, why do people accept false information as true?

According to SDT, there are two potential reasons why people show false alarms: (a) low discrimination sensitivity and (b) low response threshold. First, people may show false alarms because they are unable to accurately distinguish signal from noise. For example, in studies asking participants to judge the veracity of true and false information, participants may incorrectly identify false information as true because they are unable to accurately distinguish between true and false information. In terms of SDT, such cases can be described as instances of low discrimination sensitivity in responses to the two kinds of stimuli (see Batailler et al., 2022). Second, people may show false alarms because they tend to respond *yes, this stimulus fits the focal parameters* regardless of whether the stimulus actually fits those parameters. For example, in studies asking participants to judge the veracity of true and false information, participants may respond *true* for all information regardless of its veracity. In terms of SDT, such cases can be described as involving a low response threshold in responses to the two kinds of stimuli (see Batailler et al., 2022).

Within SDT, discrimination sensitivity and response threshold can be quantified via two distinct indices. SDT's index for discrimination sensitivity (labeled d') reflects the distance between the distributions of judgments about two stimulus classes (e.g., true vs. false information) along the judgment-relevant dimension (see Figure 1). Mathematically, discrimination sensitivity is captured by the difference between a participant's hit rate (H) and false-alarm rate (FA), with both H and FA being transformed to a quantile function for a z distribution (or inverse cumulative distribution function) in a manner such that a proportion of 0.5 is converted to a z-score of 0 (reflecting chance responses):

$$d' = z(\mathrm{H}) - z(\mathrm{FA}) \quad (1)$$

SDT's index for response threshold (labeled c) reflects the threshold along the judgment-relevant dimension at which a participant decides to switch their decision. For example, when judging

[1] In the current work, we conceptualize ideology-congruence as evaluative mismatch between new information and a person's broader network of ideological beliefs that subsumes favorable beliefs about individuals affiliated with one's own party and unfavorable beliefs about individuals affiliated with the other party (see Brandt & Sleegers, 2021).

Table 1

Signal Detection Theory Uses Hit and False-Alarm Rates to Compute d′, a Discrimination Sensitivity Index Reflecting People's Ability in Distinguishing Target Stimuli (e.g., True Information) From Distracter Stimuli (e.g., False Information), and c*, a Response Threshold Index Reflecting the Threshold for Judging Stimuli as Belonging to the Category of Target Stimuli*

Stimulus type	Response "Target" (e.g., response "true")	Response "Distracter" (e.g., response "false")
Target stimuli (e.g., true information)	Hit	Miss
Distracter stimuli (e.g., false information)	False alarm	Correct rejection

information as true versus false, *c* indicates the degree of veracity one must perceive before judging information as true (see Figure 2). Any stimulus with greater perceived veracity than that value will be judged as true, whereas any stimulus with lower perceived veracity than that value will be judged as false. In the current example, a higher (or more conservative) threshold would indicate that a participant is generally less likely to judge information as true, whereas a lower (or more liberal) threshold would indicate that a participant is generally more likely to judge information as true. Within SDT, response threshold is captured by the following equation:

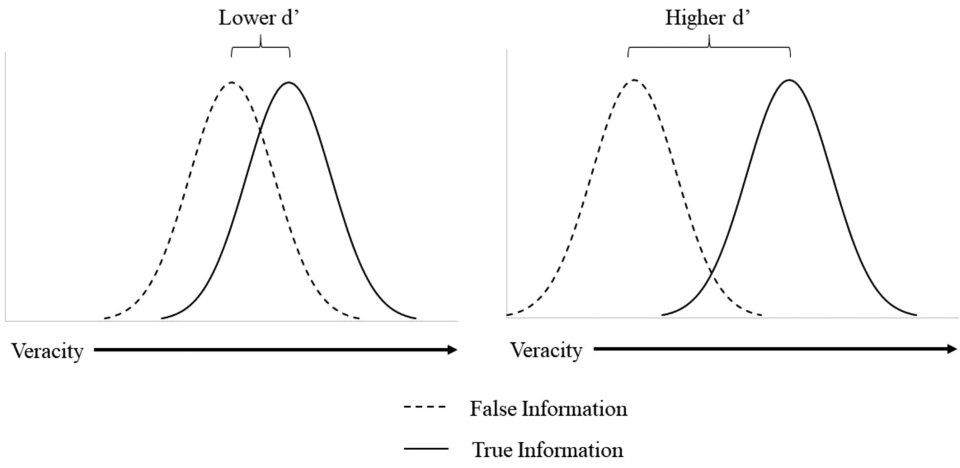
$$c = -0.5 \times [z(\mathrm{H}) + z(\mathrm{FA})] \quad (2)$$

Although *d′* and *c* are both based on hits and false alarms, the two indices are independent from one another, which means that any given factor can influence either *d′* or *c*, or both (see Macmillan & Creelman, 2004; Stanislav & Todorov, 1999). This aspect is important, because a closer examination of the debate between cognitive and motivational accounts of misinformation susceptibility reveals that their predictions are complementary rather than mutually exclusive. Similar to earlier conceptualizations emphasizing truth discernment (e.g., Pennycook & Rand, 2021), an analysis in terms of SDT suggests that the hypothesized effect of cognitive reflection should be evident in discrimination sensitivity scores (*d′*), in that greater cognitive reflection should be associated with a greater ability to distinguish between true and false information (Batailler et al., 2022). Yet, different from earlier conceptualizations equating partisan bias with lower truth discernment for ideology-congruent than ideology-incongruent information (e.g., Pennycook & Rand, 2021), an analysis in terms of SDT suggests that partisan bias should be evident in response threshold scores (*c*), in that people should show a lower threshold for judging information as true when it is congruent than when it is incongruent with their ideological beliefs (Batailler et al., 2022). Based on these considerations, we conceptualize *truth sensitivity* as accurate discrimination between true and false information, and *partisan bias* as the tendency to show a lower acceptance threshold for ideology-congruent information compared to ideology-incongruent information.

Preliminary evidence for the value of SDT in providing more nuanced insights into the underpinnings of misinformation susceptibility was provided by Batailler et al. (2022). Using SDT to reanalyze data by Pennycook and Rand (2019), Batailler et al. found that higher scores on the Cognitive Reflection Test (Frederick, 2005)—a performance-based measure claimed to capture individual differences in cognitive reflection—were associated with greater truth sensitivity in judgments of real and fake news. This result is consistent with the assumptions of cognitive accounts, suggesting that cognitive reflection increases people's ability to accurately distinguish between true and false information. In addition, Batailler et al. (2022) found a large difference in response thresholds, indicating that participants showed a much lower acceptance threshold for ideology-congruent

Figure 1

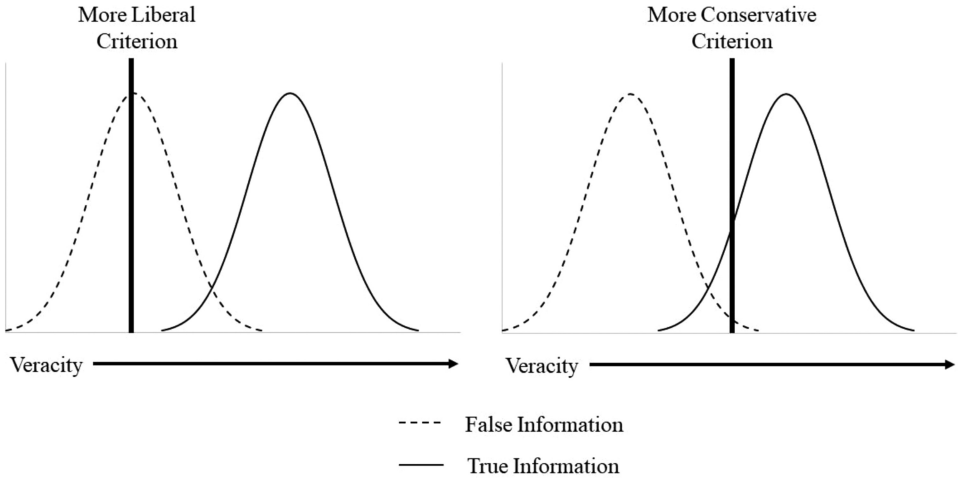
Graphical Depiction of Signal Detection Theory's Index for Discrimination Sensitivity (d′), Reflecting the Distance Between the Distributions of Judgments About True and False Information Along the Judgmental Dimension of Veracity



Note. Distributions that are closer together along the judgment-relevant dimension have a lower *d′*, indicating that participants' ability in correctly discriminating between true and false information is relatively low (left panel). Distributions that are further apart along the perceived veracity dimension have a higher *d′*, indicating that participants' ability in correctly discriminating between true and false information is relatively high (right panel).

Figure 2

Graphical Depiction of Signal Detection Theory's Index for Response Threshold (c), Reflecting the Threshold Along the Judgmental Dimension of Perceived Veracity at Which a Participant Decides to Switch Their Decision



Note. When judging whether information is true (vs. false), *c* indicates the degree of veracity the participant must perceive before judging information as true. Any stimulus with greater perceived veracity than that value will be judged as true, whereas any stimulus with lower perceived veracity than that value will be judged as false. A higher (or more conservative) threshold would indicate that a participant is generally less likely to judge information as true, whereas a lower (or more liberal) threshold would indicate that a participant is generally more likely to judge information as true.

than ideology-incongruent news. This result is consistent with the assumptions of motivational accounts, suggesting that people readily accept ideology-congruent information and reject ideology-incongruent information irrespective of the actual veracity of the respective information. Together, these findings provide preliminary evidence for the idea that susceptibility to misinformation can result from either low truth sensitivity or high partisan bias (or both).

The Current Research

Using SDT to quantify truth sensitivity and partisan bias in responses to political misinformation, the overarching goal of the current research was to gain deeper insights into the underpinnings of misinformation susceptibility. Expanding on Batailler et al.'s (2022) work, we aimed to address four issues.

First, we aimed to identify commonalities and differences between (a) judgments about the veracity of (mis)information and (b) decisions to share (mis)information online. For veracity judgments, truth sensitivity reflects the tendency to judge true information as true and false information as false, while partisan bias reflects the tendency to show a lower threshold for judging ideology-congruent information as true than ideology-incongruent information. For sharing decisions, truth sensitivity reflects the tendency to share true information and not to share false information, while partisan bias reflects the tendency to show a lower threshold for sharing ideology-congruent information than ideology-incongruent information. Although Batailler et al.'s (2022) SDT analysis of Pennycook and Rand's (2019) data provides preliminary insights into the roles of truth sensitivity and partisan bias in veracity judgments, their analysis does not include any data on sharing decisions. Because several studies have found differences between veracity judgments and sharing decisions (e.g., Pennycook, Epstein, et al., 2021; Pennycook et al., 2020), it remains unclear whether Batailler et al.'s (2022) SDT results generalize to sharing decisions. In the current research, we aimed to address this question by directly comparing veracity judgments and sharing decisions, using SDT to quantify truth sensitivity and partisan bias in either type of judgment.

Second, we aimed to identify unique and shared determinants of truth sensitivity and partisan bias in responses to misinformation. Because SDT's indices for discrimination sensitivity and response threshold are independent from one another, factors that influence truth sensitivity may not necessarily affect partisan bias, and vice versa. Preliminary evidence for this idea comes from Batailler et al.'s (2022) analysis suggesting that, although individuals with a greater propensity to engage in cognitive reflection showed higher levels of truth sensitivity, individual differences in cognitive reflection were unrelated to partisan bias. In the current research, we aimed to investigate the determinants of truth sensitivity and partisan bias more systematically, focusing on experimentally manipulated levels of cognitive reflection during encoding, self-affirmation versus self-threat, and effects of truth prompts.

Third, different from the dominant focus on fake news in this area, the current research investigated responses to misinformation more broadly. Although fake news represents an important subset of misinformation, fake news is unique in the sense that it involves "fabricated information that mimics news media content in form but […] lack(s) the news media's editorial norms and processes for ensuring the accuracy and credibility of information" (Lazer et al., 2018, p. 1094). This characteristic has led to a common confound in studies on fake-news beliefs, in that false statements are presented with a dubious news source and true statements are presented with a mainstream news source (for two notable exceptions, see Pehlivanoglu et al., 2021; Traberg & van der Linden, 2022). Although this confound

may be common for real and fake news encountered in natural contexts, it creates ambiguity about whether a given finding is driven by the veracity of the statements or the trustworthiness of the source. For example, Batailler et al.'s (2022) finding that individuals high (vs. low) in cognitive reflection show a greater sensitivity in distinguishing between real and fake news may be driven by (a) individual differences in the ability to distinguish between true and false statements or (b) individual differences in the use of information about the source's trustworthiness. This difference is important because fake news constitutes only a very small fraction of all misinformation that is spreading online (see Grinberg et al., 2019; Guess et al., 2019; Nelson & Taneja, 2018). In fact, most misinformation does not qualify as fake news in terms of the scientific definition (see Lazer et al., 2018) because it comes from sources that do not pretend to be news media. Thus, to understand why people fall for misinformation more broadly (rather than fake news in particular), it is important to investigate responses to misinformation without confounding statement veracity and source reliability. In the current studies, we aimed to address this concern by investigating responses to true and false statements in the absence of information about their source.

Fourth, we investigated the relative extent to which susceptibility to misinformation is accounted for by truth sensitivity and partisan bias, respectively. From the perspective of SDT, the question of why people fall for misinformation can be understood as being concerned with the causes of false alarms (see Table 1). For veracity judgments, this question translates into: Why do people believe false information? For sharing decisions, this question translates into: Why do people share false information? Truth sensitivity and partisan bias can be understood as two distinct factors that may independently predict the extent to which people believe in or share false information. In the current studies, we aimed to gain deeper insights into the underpinnings of misinformation susceptibility by examining the relative extent to which belief in and sharing of false information is accounted for by truth sensitivity and partisan bias.

Toward these ends, we conducted four preregistered experiments, using SDT to quantify truth sensitivity and partisan bias in responses to political misinformation. Experiment 1 compared the degrees of truth sensitivity and partisan bias in veracity judgments and sharing decisions. Experiment 2 tested effects of cognitive reflection during encoding on truth sensitivity and partisan bias in veracity judgments and sharing decisions. Experiment 3 aimed to provide deeper insights into the determinants of partisan bias by investigating effects of self-affirmation versus self-threat on partisan bias in veracity judgments and sharing decisions. Experiment 4 tested effects of truth prompts on truth sensitivity and partisan bias in sharing decisions. Expanding on the findings of the four individual studies, we also conducted integrative analyses of the combined data (see Curran & Hussong, 2009) to investigate the relative extent to which belief in and sharing of false information is accounted for by truth sensitivity and partisan bias, respectively.

Transparency and Openness

For each study, we report how we determined our sample size, all data exclusions, all manipulations, and all measures. The data, analysis codes, and research materials of all studies are available at https://osf.io/d2rne/. The Open Science Framework (OSF) page also includes a brief tutorial on the use of SDT in research on misinformation. Data were analyzed using PASW Statistics 18. All power analyses were conducted using GPower 3.1.9.7 (Faul et al., 2007). The design, hypotheses, and analysis plan were pre-registe red for all four experiments. Hyperlinks to the individual pre-registrations are provided in the Method section of each study. The re ported research received ethical approval from the Institutional Re view Board of the University of Texas at Austin.

Experiment 1

Experiment 1 directly compared veracity judgments and sharing decisions in terms of truth sensitivity and partisan bias. There are several reasons why veracity judgments and sharing decisions may diverge in terms of truth sensitivity and partisan bias. On the one hand, there is evidence suggesting that people often pay insufficient attention to veracity when they share information online (e.g., Pennycook, Epstein, et al., 2021; Pennycook et al., 2020). In SDT terms, such cases should be reflected in a lower degree of truth sensitivity for sharing decisions compared to veracity judgments. On the other hand, it seems plausible that people do not share all information they know is true, which seems most likely for information that conflicts with their personal beliefs. In SDT terms, these asymmetries should be reflected in greater levels of partisan bias in sharing decisions compared to veracity judgments. Based on these considerations, Experiment 1 served as an initial test of whether veracity judgments and sharing decisions differ in terms of truth sensitivity and partisan bias.

Toward this end, self-identified Republicans and self-identified Democrats were presented with a series of news headlines gathered from the Internet, half of which were true and half of which were false. Orthogonal to the manipulation of veracity, the headlines were selected such that half of them had a pro-Republican slant whereas the other half had a pro-Democrat slant (confirmed in a pilot study prior to Experiment 1). To investigate potential differences between veracity judgments and sharing decisions, half of the participants were asked to indicate for each headline if, to the best of their knowledge, it is true or false; the remaining half were asked to indicate for each headline whether they would share the story online (see Pennycook & Rand, 2019). Responses were analyzed using SDT to quantify the degree of truth sensitivity and partisan bias in veracity judgments and sharing decisions, respectively. Using SDT's indices for discrimination sensitivity (d') and response threshold (c), Experiment 1 tested the following two preregistered hypotheses:

> *Hypothesis 1:* Truth sensitivity (captured by SDT's d' index) will significantly differ for veracity judgments versus sharing decisions.

> *Hypothesis 2:* Partisan bias (captured by the difference between SDT's c index for ideology-congruent and ideology-incongruent news headlines) will significantly differ for veracity judgments versus sharing decisions.

Expanding on the tests of our preregistered hypotheses, we also conducted exploratory analyses using a measure of self-perceived ability in identifying news that is made up. Using the same instrument, Lyons et al. (2021) obtained evidence for a better-than-average effect (see Alicke & Govorun, 2005), in that most participants rated themselves above average in their ability in identifying made-up news. In addition, the authors obtained evidence

for a meta-ignorance effect (a.k.a. Dunning–Kruger effect; see Dunning, 2011), in that even participants who showed the lowest accuracy levels in distinguishing between true and false news headlines rated themselves as above average in the ability to identify made-up news. This effect was observed for both judgments of veracity and willingness to share information on social media. In the current study, we aimed to replicate Lyons et al.'s (2021) findings for truth sensitivity and further explored relations of self-perceived ability in identifying made-up news with truth sensitivity partisan bias, respectively. Toward this end, we conducted exploratory analyses investigating correlations between self-perceived ability in identifying made-up news and truth sensitivity in veracity judgments and sharing decisions, respectively. Correspondingly, we conducted exploratory analyses investigating correlations between self-perceived ability in identifying made-up news and partisan bias in veracity judgments and sharing decisions, respectively.

Method

Preregistration

The design, hypotheses, and analysis plan of Experiment 1 were preregistered prior to data collection at https://osf.io/ud5j8/. The data for Experiment 1 were collected in June 2021.

Participants and Design

We aimed to have at least 80% power for the detection of a small between-group difference of $d = 0.30$ in a t-test for independent means with an α-level of 0.05 (two-tailed), which requires a sample of 352 participants. For the critical tests in the current study (see below), a sample of this size provides a power of 80% in detecting a small effect of $f = 0.089$ in a 2×2 mixed analysis of variance (ANOVA) with one factor varying between-subjects and the other varying within-subjects (two-tailed), assuming a correlation of $r = .30$ between measures and using a nonsphericity correction of $\varepsilon = 1$. Anticipating that approximately 10% of the participants would fail to pass our attention check (see below), we set our preregistered target sample to 400 participants prior to exclusions.

Participants were recruited on Prolific Academic, a crowdsourcing platform that provides access to diverse samples of participants for psychological online research (Peer et al., 2017). To obtain a balanced sample of participants who identify as either Democrat or Republican, we created two assignments for the separate recruitment of 200 self-identified Democrats and 200 self-identified Republicans. To this end, we used Prolific's prescreening filters to restrict completion of one assignment to participants who self-identify as Democrat and completion of the other assignment to participants who self-identify as Republican. For both assignments, additional preregistered filters were used to restrict participation to Prolific workers who (a) are 18 years old or older, (b) have an approval rating of >95% on prior assignments on Prolific, (c) are a citizen of the United States, (d) are a resident of the United States, (e) have completed at least 100 prior assignments on Prolific, and (f) are fluent in English. The study took approximately 10–15 min to complete, and participants were compensated $3.00 for their time.

Following our preregistered stopping rule, data collection ended once 400 Prolific workers had been approved for credit. Of the 412 Prolific workers who started the study, 400 completed all measures. Of the 400 participants with complete data, 25 failed to pass an attention check and two reported inconsistent political affiliations in Prolific's prescreening survey and the measure of political affiliation included in the current study. In line with our preregistered exclusion criteria (see below), data from these participants were excluded from analyses. Thus, the final sample included a total of 373 participants (150 men, 213 women, one prefer not to answer, nine other), 189 of which identified as Democrat and 184 of which identified as Republican ($n = 185$ in the veracity-judgment condition; $n = 188$ in the sharing-decision condition).[2] Participants' age ranged from 18 to 70 years ($M_{age} = 33.35$ years, $SD_{age} = 11.36$). Of the 373 participants in the final sample, 301 identified as White, 21 identified as Black or African American, 3 identified as American Indian or Alaska Native, 35 identified as Asian, zero identified as Native Hawaiian or Pacific Islander, four as other, and nine identified with more than one race category. For household income, 38 reported incomes lower than $20,000, 71 reported incomes between $20,000 and $40,000, 77 reported incomes between $40,000 and $60,000, 52 reported incomes between $60,000 and $80,000, 46 reported incomes between $80,000 and $100,000, and 89 reported incomes higher than $100,000. For education, 0 reported having less than a high school degree, 42 reported having a high school degree or equivalent, 86 reported having some college education with no degree, 31 reported having a 2-year college associate degree, 151 reported having a 4-year college bachelor's degree, 51 reported having a master's degree, five reported having a doctoral degree, and seven reported having a professional JD or MD degree. Self-identified Republicans considered themselves significantly more conservative than self-identified Democrats in general (Ms = 5.63 vs. 1.74), $t(327.98) = 38.72$, $p < .001$, $d = 4.01$, in terms of economic issues (Ms = 5.75 vs. 1.94), $t(357.65) = 32.16$, $p < .001$, $d = 3.33$, and in terms of social issues (Ms = 5.41 vs. 1.56), $t(285.07) = 33.57$, $p < .001$, $d = 3.48$.

The study included a 2 (Headline Accuracy: true vs. false) × 2 (Headline Slant: pro-Democrat vs. pro-Republican) × 2 (Political Affiliation: Democrat vs. Republican) × 2 (Judgment Type: veracity-judgment vs. sharing-decision) mixed design with the first two factors varying within-subjects and the latter two factors varying between-subjects. Participants were randomly assigned to either the veracity-judgment or the sharing-decision condition.

Procedure and Materials

Participants read a set of 60 news headlines that varied in terms of whether (a) the claim in the headline is true or false and (b) the headline has a pro-Democrat or a pro-Republican slant (15 headlines per category). The 60 headlines were selected from a larger pool of headlines gathered from the Internet, thoroughly prescreened in terms of basic criteria, and subjected to a pilot study to confirm their essential properties (see Pennycook, Binnendyk, et al., 2021). The procedure for the selection of headlines is

[2] Exploratory analyses investigating potential differences between self-identified Democrats and self-identified Republicans are presented in the online supplemental materials.

described in the Appendix. The results of the pilot study and the final list of headlines used in the current study are provided at https://osf.io/d2rne/. Participants were told that they will be shown a series of politically relevant statements gathered on the Internet, instructed to read each statement carefully, and asked to answer a question about each headline. In the veracity-judgment condition, headlines were presented one at a time together with the question *To the best of your knowledge, is the claim in this headline true or false?* and the response options *true* and *false*. In the sharing-decision condition, headlines were presented one at a time together with the question *Would you consider sharing this story online (e.g., through social media or other platforms)?* and the response options *yes* and *no*. The order of the headlines was randomized individually for each participant. After responding to the headlines, participants were asked to complete measures assessing their political ideology, political affiliation, and self-perceived ability to identify news that is made up. Participants then completed a set of demographic questions and an attention check. Finally, participants were debriefed, thanked for their participation, and redirected for compensation.

Supplemental Measures

Political Ideology. We assessed participants' political ideology using three items: (a) *How do you consider yourself politically in general?* (b) *How do you consider yourself politically in terms of economic issues?* (c) *How do you consider yourself politically in terms of social issues?* Responses were measured with 7-point rating scales ranging from 1 (*very liberal*) to 7 (*very conservative*).

Political Affiliation. To confirm participants' self-reported political affiliation in Prolific's prescreening survey, participants were asked to indicate if they think of themselves as Democrat or Republican. To this end, participants were presented with the question *Generally speaking, do you usually think of yourself as a Republican or a Democrat?* and the response options *Republican* and *Democrat*.

Self-Perceived Ability. For exploratory purposes, the study included two questions adapted from Lyons et al. (2021) asking participants to rate their ability in identifying news that is made up. The first question was: *How do you think you compare to other Americans in your general ability to recognize news that is made up? Please respond using the scale below, where 1 means you are at the very bottom (worse than 99% of people) and 100 means you are at the very top (better than 99% of people).* The second question was: *How do you think you compare to other Americans in how well you performed in this study at recognizing news that is made up? Please respond using the scale below, where 1 means you are at the very bottom (worse than 99% of people) and 100 means you are at the very top (better than 99% of people).* Responses to both questions were measured with a sliding scale ranging from 0 to 100.

Demographics. Participant gender was measured with the question *What is your gender?* and the four response options (a) *male*, (b) *female*, (c) *prefer not to answer*, and (d) *other* with a textbox for further specification. Age was measured with the question *What is your age?* and a textbox for numeric responses. Ethnicity was measured with the question *Are you Spanish, Hispanic, or Latino or none of these?* and the two response options (a) *yes* and (b) *none of these*. Race was measured with the question *Choose one or more races that you consider yourself to be* and the six response options (a) *White*, (b) *Black or African American*, (c) *American Indian or Alaska Native*, (d) *Asian*, (e) *Native Hawaiian or Pacific Islander*, and (f) *Other* with a textbox for further specification. Education level was measured with the question *What is the highest level of school you have completed or the highest degree you have received?* and the eight response options (a) *Less than high school degree*, (b) *High school graduate (high school diploma or equivalent including GED)*, (c) *Some college but no degree*, (d) *Associate degree in college (2-year)*, (e) *Bachelor's degree in college (4-year)*, (f) *Master's degree*, (g) *Doctoral degree*, and (h) *Professional degree (JD, MD)*. Income was measured with the question *What is your annual household income?* and the six response options (a) *Under $20,000*, (b) *$20,001–$40,000*, (c) *$40,001–$60,000*, (d) *$60,001–$80,000*, (e) *$80,001–$100,000*, and (f) *$100,000+*.

Attention Check. A reading-intensive attention check was used to identify inattentive participants (Oppenheimer et al., 2009). The attention check required participants to read a set of instructions that asked participants to not answer a question. The instructions were as follows:

> To facilitate our research on decision-making we are interested in learning a little more about you, the decision-maker. Psychological research using text-based materials requires that study participants read the materials and do not skip over longer pieces of text. We are therefore interested in whether you actually take the time to read the directions; if not, then some of our manipulations that rely on changes in the instructions will be ineffective. To demonstrate that you have read the instructions, please ignore the question below and all of the response options. Instead, simply continue on to the next page without answering the question. Thank you very much.

Following the instructions, participants were presented with the question *Of the following destinations, which one would be your first choice for a vacation if you had a free all-inclusive round trip after the Covid-19 pandemic? (Check all that apply)* and the response options *Australia, Brazil, China, Egypt, France, Germany, India, Japan, New Zealand, Mexico, Russia, South Africa, Spain, Sweden*, and *United Kingdom*. Because the instructions directed participants to not select any options and instead skip ahead to the next screen, those who checked one or more of the 15 response options were considered to have failed the attention check.

Data Aggregation and Treatment

Hit Rates. Following our preregistered data aggregation plan, hit rates (H) were calculated as the proportion of true news headlines judged as true (in the veracity-judgment condition) or the proportion of true news headlines that participants would share (in the sharing-decision condition). Hit rates were calculated separately for headlines that are congruent with participants' political affiliation (i.e., pro-Democrat headlines for self-identified Democrats; pro-Republican headlines for self-identified Republicans) and for headlines that are incongruent with participants' political affiliation (i.e., pro-Republican headlines for self-identified Democrats; pro-Democrat headlines for self-identified Republicans).

False-Alarm Rates. Following our preregistered data aggregation plan, false-alarm rates (FA) were calculated as the proportion of false news headlines judged as true (in the veracity-judgment

condition) or the proportion of false news headlines that participants would share (in the sharing-decisions condition). False-alarm rates were calculated separately for headlines that are congruent with participants' political affiliation (i.e., pro-Democrat headlines for self-identified Democrats; pro-Republican headlines for self-identified Republicans) and for headlines that are incongruent with participants' political affiliation (i.e., pro-Republican headlines for self-identified Democrats; pro-Democrat headlines for self-identified Republicans).

SDT Indices. Following our preregistered data aggregation plan, we used SDT to calculate d' scores for each participant, which can be interpreted as an index of truth sensitivity. Scores were calculated separately for ideology-congruent and ideology-incongruent headlines using the equation: $d' = z(H) - z(FA)$. Scores were calculated such that higher values reflect greater sensitivity to actual information veracity in veracity judgments and sharing decisions, respectively. In addition, we calculated c scores for each participant, reflecting the threshold along the truth dimension at which a participant decides to switch their decision. Response threshold scores were calculated separately for ideology-congruent and ideology-incongruent headlines using the equation: $c = -0.5 \times [z(H) + z(FA)]$. Scores were calculated such that lower values reflect a lower acceptance threshold for judging headlines as true or for decisions to share headlines. The difference in c scores for ideology-incongruent and ideology-congruent headlines was interpreted as an indicator of partisan bias, conceptualized as lower threshold for accepting ideology-congruent than ideology-incongruent headlines.

Deviation From Preregistered Data Aggregation Plan. All indices were calculated in line with our preregistered data aggregation plan with one exception. In cases where the proportion of *true* or *yes* responses within a given headline-category is either 0 or 1, it is not possible to calculate d' and c scores using the standard SDT formulas. In such cases, we followed recommendations by MacMillan and Creelman (2004) and converted values of 0 to $1/(2N)$ and values of 1 to $1-1/(2N)$, where N is the number of trials (i.e., 15 in our case). We did not anticipate this issue prior to analyzing the data of Experiment 1 and therefore did not preregister this transformation. The transformation was included in the preregistrations for all subsequent studies.

Missing Data and Data Exclusions. The survey was programmed such that all questions required a response. As a result, there were no missing data except from participants who did not complete the study until the end. Following our preregistered analysis plan, we excluded data from participants who did not complete the entire study, failed the attention check, or showed inconsistent self-reports of their political affiliations in Prolific's prescreening and the measure of political affiliation in the current study (i.e., we included only those participants whose self-reported political affiliation in the demographic measure was identical to the one they reported in Prolific's prescreening).

Results

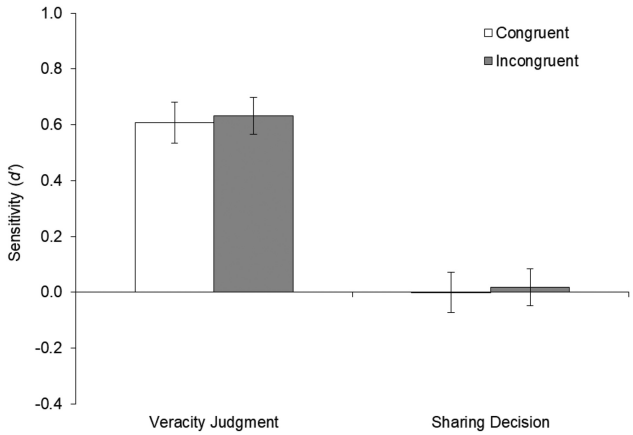
Confirmatory Analyses

Truth Sensitivity. To investigate potential differences in truth sensitivity as a function of Judgment Type, we preregistered that d' scores will be submitted to a 2 (Judgment Type: veracity-judgment vs. sharing-decision) × 2 (Ideology-Congruence: congruent vs. incongruent) ANOVA with the first variable as a between-subjects factor and the second variable as a within-subject factor. Means and 95% confidence intervals in the four conditions are presented in Figure 3. Confirming Hypothesis 1, the ANOVA revealed a significant main effect of Judgment Type, $F(1, 371) = 240.04$, $p < .001$, $\eta_p^2 = 0.393$, indicating that truth sensitivity was greater for veracity judgments than sharing decisions. The main effect of Ideology-Congruence and the two-way interaction between Judgment Type and Ideology-Congruence were not statistically significant (all Fs < 1, all ps $> .46$).

Response Threshold. To investigate potential differences in partisan bias as a function of Judgment Type, we preregistered that c scores will be submitted to a 2 (Judgment Type: truth-judgment vs. sharing-decision) × 2 (Ideology-Congruence: congruent vs. incongruent) ANOVA with the first variable as a between-subjects factor and the second as a within-subject factor. Means and 95% confidence intervals in the four conditions are presented in Figure 4. Confirming the presence of partisan bias, the ANOVA revealed a significant main effect of Ideology-Congruence, $F(1, 371) = 382.06$, $p < .001$, $\eta_p^2 = 0.507$, indicating that participants showed a lower acceptance threshold for ideology-congruent headlines compared to ideology-incongruent headlines. A significant main effect of Judgment Type further revealed that participants had a higher acceptance threshold for sharing decisions compared to veracity judgments, $F(1, 371) = 179.56$, $p < .001$, $\eta_p^2 = 0.326$. Counter to Hypothesis 2, the two-way interaction of Judgment Type and Ideology-Congruence was not statistically significant, $F(1, 371) = 0.71$, $p = .400$, $\eta_p^2 = 0.002$. A significant partisan-bias effect emerged for both veracity judgments, $F(1, 184) = 260.28$, $p < .001$, $\eta_p^2 = 0.586$, and sharing decisions, $F(1, 187) = 157.79$, $p < .001$, $\eta_p^2 = 0.458$, as revealed by significant main effects of Ideology-Congruence within each of the two Judgment Type conditions.

Figure 3

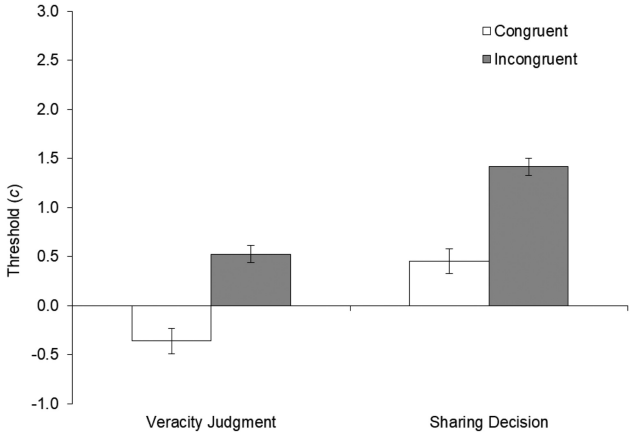
Signal Detection d' Scores Reflecting Truth Sensitivity in Responses to Political Information as a Function of Ideology-Congruence (Congruent vs. Incongruent) and Judgment Type (Veracity Judgment vs. Sharing Decision), Experiment 1



Note. Higher d' scores reflect greater truth sensitivity. Error bars depict 95% confidence intervals.

Figure 4

Signal Detection c *Scores Reflecting Response Threshold in Responses to Political Information as a Function of Ideology-Congruence (Congruent vs. Incongruent) and Judgment Type (Veracity Judgment vs. Sharing Decision), Experiment 1*



Note. Higher *c* scores reflect higher acceptance threshold. Error bars depict 95% confidence intervals.

Exploratory Analyses

Self-Perceived Ability. Responses to the two items measuring self-perceived ability in identifying made-up news were combined in a single index by averaging responses to the two items (Cronbach's $\alpha = 0.83$). In addition, we created a partisan-bias index by calculating the difference in *c* scores for ideology-congruent and ideology-incongruent news, such that higher scores reflect a lower threshold for accepting ideology-congruent headlines compared to ideology-incongruent headlines (i.e., greater partisan bias). An index of overall truth sensitivity was created by averaging d' scores for ideology-congruent and ideology-incongruent headlines. Overall truth sensitivity and partisan bias were not significantly correlated for veracity judgments ($r = .139$, $p = .059$) and sharing decisions ($r = .046$, $p = .532$).

Replicating findings by Lyons et al. (2021), participants showed a better-than-average effect in self-perceived ability ratings, in that 76.9% of the participants rated themselves as above average in their ability to recognize made-up news. Mean ratings were significantly greater than 50% with a self-perceived ability score of 62.47%, $t(372) = 13.64$, $p < .001$, $d = 0.706$. For veracity judgments, self-perceived ability in identifying made-up news showed a significant positive correlation with overall truth sensitivity ($r = .265$, $p < .001$), indicating that participants who perceived themselves as better at identifying made-up news were indeed better in distinguishing between true and false headlines. A positive correlation between self-perceived ability in identifying made-up news and partisan bias ($r = .248$, $p = .001$) further suggested that participants who perceived themselves as better at identifying made-up news showed a greater partisan bias in their veracity judgments. For sharing decisions, self-perceived ability in identifying made-up news was not significantly correlated with overall truth sensitivity ($r = .054$, $p = .458$) or partisan bias ($r = .104$, $p = .156$).

Discussion

Experiment 1 revealed three sets of noteworthy findings. First, consistent with Hypothesis 1, our preregistered confirmatory analyses revealed that truth sensitivity was greater for veracity judgments than sharing decisions. The effect size of this difference qualifies as large in terms of current conventions (J. Cohen, 1988), suggesting that actual information veracity had a much greater impact on veracity judgments than sharing decisions. In fact, the 95% confidence intervals for truth sensitivity in sharing decisions included zero for both ideology-congruent and ideology-incongruent information (see Figure 3), indicating that for either type of information participants were as likely to share false information as they were to share true information. Put differently, although participants showed a considerable ability in distinguishing between true and false headlines when they were asked to judge the veracity of the headlines, sharing decisions were completely unaffected by actual information veracity.

Second, participants' acceptance threshold was higher for sharing decisions than veracity judgments. The effect size of this difference also qualifies as large in terms of current conventions (J. Cohen, 1988), suggesting that participants were much more reluctant in sharing information than accepting information as true. Interestingly, participants' reluctance in sharing information was not associated with greater truth sensitivity, as indicated by the lack of truth sensitivity in sharing decisions (see above). We also obtained evidence for a large partisan-bias effect, in that participants showed a substantially lower threshold for accepting ideology-congruent than ideology-incongruent information. The effect size of this difference also qualifies as large in terms of current conventions (J. Cohen, 1988). Yet, counter to Hypothesis 2, partisan bias did not significantly differ across judgment types, in that partisan bias was similarly large for veracity judgments and sharing decisions. Thus, counter to the idea that partisan bias would be more pronounced for sharing decisions than veracity judgments, it seems that partisan bias plays an equally important role for what people believe to be true and what people share online.

Third, replicating prior findings by Lyons et al. (2021), exploratory analyses revealed evidence for a better-than-average effect in participants' self-perceived ability to recognize made-up news, in that more than 75% of the participants rated themselves as above average in this particular ability. Moreover, although participants who rated themselves higher in the ability to identify made-up news showed higher levels of truth sensitivity in veracity judgments, self-perceived ability also showed a positive correlation with partisan bias in veracity judgments, in that those who were more confident about their ability showed greater partisan bias. Self-perceived ability to recognize made-up news was unrelated to truth sensitivity and partisan bias in sharing decisions. We will return to these findings in Experiment 3 when we discuss potential underpinnings of partisan bias.

Experiment 2

The main goal of Experiment 2 was to investigate effects of cognitive reflection on truth sensitivity and partisan bias in veracity

judgments and sharing decisions. Previous research suggests that greater cognitive reflection is associated with greater truth discernment in responses to misinformation (for a review, see Pennycook & Rand, 2021). However, the majority of these studies have relied on correlational approaches to investigate associations between truth discernment and individual differences in cognitive reflection (e.g., Mosleh et al., 2021; Pennycook & Rand, 2019; Ross et al., 2021). Although it is possible that the obtained associations reflect a proximal effect of cognitive reflection during the encoding of true and false information, they could also reflect a distal effect where the obtained associations are driven by a factor that is independent of cognitive reflection during the encoding of true and false information. For example, individuals high in cognitive reflection may be more likely to seek fact-checking information in their daily routines, which, in turn, may increase their knowledge base for judgments of true and false information. In that case, the critical factor underlying the obtained correlations would be topic-relevant knowledge levels, not cognitive reflection during the encoding of true and false information. The difference between the two possibilities is important for both theoretical and practical reasons because nudges to engage in more elaborate processing during the encoding of true and false information should increase truth discernment only if cognitive reflection has a proximal effect, but not if it has a distal effect.

One experimental study suggests that cognitive reflection might indeed have a proximal effect on responses to misinformation, showing that truth discernment was greater under high-reflection compared to low-reflection conditions (Bago et al., 2020). In this work, however, false statements were always presented with a dubious news source and true statements were always presented with a mainstream news source—a common confound in research on fake-news beliefs (see above). It therefore remains unclear whether cognitive reflection increased truth discernment via (a) enhanced ability to distinguish between true and false statements or (b) enhanced reliance on information about the sources' trustworthiness. Based on the currently available evidence, it also remains unclear whether cognitive reflection during encoding is effective in reducing partisan bias, and whether effects of cognitive reflection on truth sensitivity and partisan bias are equivalent for veracity judgments and sharing decisions.

Experiment 2 aimed to address these questions. Toward this end, self-identified Republicans and self-identified Democrats were presented with true and false news headlines that had either a pro-Republican or pro-Democrat slant. Following the procedure in Experiment 1, half of the participants were asked to indicate for each headline if it is true or false; the remaining half were asked to indicate for each headline whether they would share the story online. Orthogonal to the manipulation of judgment type, half of the participants were asked to report their initial reaction to each headline and given 7 seconds to do so (low-reflection condition); the remaining half were asked to read the headline carefully and take a moment to think about their answer without time limits (high-reflection condition). Responses were analyzed using SDT to quantify the degree of truth sensitivity and partisan bias in veracity judgments and sharing decisions, respectively. Based on the results of prior experimental work (Bago et al., 2020) and reanalyses of existing data suggesting that individual differences in cognitive reflection are significantly related to truth sensitivity, but not partisan bias, in veracity judgments of real and fake news (Batailler et al., 2022), we tested the hypothesis that experimentally manipulated levels of cognitive reflection would show similar effects on responses to true and false statements presented without information about their source. Specifically, using SDT's indices for discrimination sensitivity (d') and response threshold (c), Experiment 2 tested the following two preregistered hypotheses:

Hypothesis 1: Cognitive reflection will increase truth sensitivity captured by SDT's d' index (Hypothesis 1a), and this increase will be observed for both veracity judgments (Hypothesis 1b) and sharing decisions (Hypothesis 1c).

Hypothesis 2: Cognitive reflection will have no effect on partisan bias captured by the difference between SDT's c index for ideology-congruent and ideology-incongruent news headlines (Hypothesis 2a), and this null effect will be observed for both veracity judgments (Hypothesis 2b) and sharing decisions (Hypothesis 2c).

Method

Preregistration

The design, hypotheses, and analysis plan of Experiment 2 were preregistered prior to data collection at https://osf.io/6y325/. The data for Experiment 2 were collected in June 2021.

Participants and Design

As one of the two preregistered hypotheses involved a null effect, we increased our desired statistical power from 80% in Experiment 1 to 95% in Experiment 2. Specifically, we aimed to have 95% power for the detection of a small effect of $f = 0.10$ in a 2 (Cognitive Reflection: low vs. high) × 2 (Judgment Type: veracity-judgment vs. sharing-decision) × 2 (Ideology-Congruence: congruent vs. incongruent) mixed ANOVA with the first two variables as between-subjects factors and the last one as a within-subject factor (two-tailed). Assuming a correlation between measures of $r = .30$ and using a nonsphericity correction of $\varepsilon = 1$, a sample size of at least 608 participants is required to meet this target. Anticipating that approximately 10% of the participants may fail to pass our attention check and to obtain sufficient power for slightly smaller effects, we set our preregistered target sample to 800 participants prior to exclusions. Participants were recruited on Prolific Academic via two assignments for the separate recruitment of 400 self-identified Democrats and 400 self-identified Republicans. Eligibility criteria for participation were identical to Experiment 1. The study took approximately 10–15 min to complete, and participants were compensated $3.00 for their time.

Following our preregistered stopping rule, data collection ended once 800 Prolific workers had been approved for credit. Of the 819 Prolific workers who started the study, 800 completed all measures. Of the 800 participants with complete data, 48 failed to pass an attention check and 10 reported inconsistent political affiliations in Prolific's prescreening survey and the current study. Data from these participants were excluded from analyses. Following our preregistered exclusion criteria, we also excluded data from 12 participants in the low-reflection condition who failed to respond within the 7-s response window for more than five headlines within one or more of the four headline categories. Thus, the final sample included a total of 730 participants (406 men, 314 women, six prefer not to answer, four other), 366 of which identified as Democrat and 364 of which identified as Republican ($n = 180$ in the veracity-

judgment/low-reflection condition; $n = 181$ in the veracity-judgment/high-reflection condition; $n = 183$ in the sharing-decision/low-reflection condition; $n = 186$ in the sharing-decision/high-reflection condition).[3] Participants' age ranged from 18 to 92 years ($M_{age} = 36.19$ years, $SD_{age} = 12.79$). Of the 730 participants in the final sample, 526 identified as White, 92 identified as Black or African American, five identified as American Indian or Alaska Native, 65 identified as Asian, one identified as Native Hawaiian or Pacific Islander, 14 as other, and 27 identified with more than one race category. For household income, 72 reported incomes lower than $20,000, 141 reported incomes between $20,000 and $40,000, 126 reported incomes between $40,000 and $60,000, 117 reported incomes between $60,000 and $80,000, 71 reported incomes between $80,000 and $100,000, and 203 reported incomes higher than $100,000. For education, three reported having less than a high school degree, 79 reported having a high school degree or equivalent, 125 reported having some college education with no degree, 63 reported having a 2-year college associate degree, 277 reported having a 4-year college bachelor's degree, 155 reported having a master's degree, 18 reported having a doctoral degree, and 10 reported having a professional JD or MD degree. Self-identified Republicans considered themselves significantly more conservative than self-identified Democrats in general (Ms = 5.61 vs. 2.06), $t(728) = 40.64$, $p < .001$, $d = 3.01$, in terms of economic issues (Ms = 5.63 vs. 2.28), $t(728) = 35.33$, $p < .001$, $d = 2.61$, and in terms of social issues (Ms = 5.39 vs. 1.89), $t(709.79) = 35.81$, $p < .001$, $d = 2.65$.

The study included a 2 (Headline Accuracy: true vs. false) × 2 (Headline Slant: pro-Democrat vs. pro-Republican) × 2 (Political Affiliation: Democrat vs. Republican) × 2 (Judgment Type: veracity-judgment vs. sharing-decision) × 2 (Cognitive Reflection: low vs. high) mixed design with the first two factors varying within-subjects and the latter three factors varying between-subjects. Participants were randomly assigned to one of the four between-subjects conditions.

Procedure, Materials, and Measures

The procedures, materials, and measures were identical to Experiment 1 with three exceptions. First, Experiment 2 did not include the exploratory measure of self-perceived ability to identify made-up news. Second, to obtain more information about our participants, we added a measure of social-media use and political interest to the demographic survey at the end. Third, to manipulate cognitive reflection, participants were asked to report their initial reaction to each statement (low-reflection condition) or to read the headline carefully and take a moment to think about their answer (high-reflection condition). Participants in the low-reflection were instructed to provide their response within 7-s, and the program moved forward to the next headline if participants did not provide a response within the 7-s response window. Participants in the high-reflection condition had unlimited time to provide their response.

Data Aggregation and Treatment

The preregistered data aggregation plan followed the procedures in Experiment 1, the only difference being that we preregistered the transformation of data for cases where the proportion of *true* or *yes* responses within a given headline-category is either 0 or 1. In such cases, we converted values of 0 to $1/(2N)$ and values of 1 to $1-1/(2N)$, where N is the number of trials within a given headline category (see MacMillan & Creelman, 2004). We applied the same exclusion criteria as Experiment 1. Additionally, in accordance with our preregistered data analysis plan, we excluded participants who failed to respond within the 7-s response window for more than five headlines within one or more of the four headline categories.

Results

Manipulation Check

To investigate the effectiveness of our manipulation of cognitive reflection, we tested whether the total time participants took to complete the study differed across cognitive-reflection conditions. Supporting the effectiveness of our cognitive-reflection manipulation, participants in the high-reflection condition took significantly more time to complete the study ($M = 835.28$ s, $SD = 676.95$) than participants in the low-reflection condition ($M = 584.15$ s, $SD = 331.98$), $t(533.72) = 6.37$, $p < .001$, $d = 0.470$.

Confirmatory Analyses

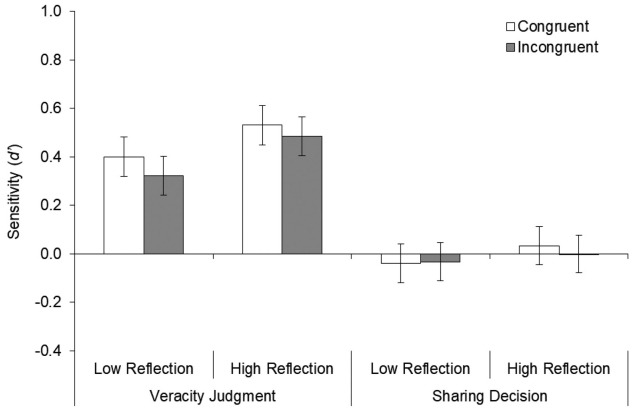
Truth Sensitivity. To investigate effects of cognitive reflection on truth sensitivity as a function of veracity judgments versus sharing decisions, d' scores were submitted to a 2 (Cognitive Reflection: low vs. high) × 2 (Judgment Type: veracity-judgment vs. sharing-decision) × 2 (Ideology-Congruence: congruent vs. incongruent) ANOVA with the first two variables as between-subjects factors and the third as a within-subject factor. Means and 95% confidence intervals across conditions are presented in Figure 5. Replicating the results of Experiment 1, the analysis revealed a significant main effect of Judgment Type, $F(1, 726) = 200.98$, $p < .001$, $\eta_p^2 = 0.217$, indicating that truth sensitivity was greater for veracity judgments than sharing decisions. Consistent with Hypothesis 1a, a significant main effect of Cognitive Reflection indicated that truth sensitivity was greater in the high-reflection condition than the low reflection condition, $F(1, 726) = 9.81$, $p = .002$, $\eta_p^2 = 0.013$. Moreover, consistent with Hypothesis 1b, the main effect of Cognitive Reflection was statistically significant for veracity judgments, $F(1, 359) = 7.91$, $p = .005$, $\eta_p^2 = 0.022$. However, inconsistent with Hypothesis 1c, the main effect of Cognitive Reflection was not statistically significant for sharing decisions, $F(1, 367) = 2.02$, $p = .156$, $\eta_p^2 = 0.005$. Although the latter findings may suggest that Cognitive Reflection influences veracity judgments but not sharing decisions, such a conclusion conflicts with the lack of a significant interaction between Cognitive Reflection and Judgment Type (all Fs < 2.28, all ps > .13). Thus, although the current findings provide clear support for Hypotheses 1a and 1b, the current data are inconclusive regarding Hypothesis 1c.

Response Threshold. To investigate effects of cognitive reflection on partisan bias as a function of veracity judgments versus sharing decisions, c scores were submitted to a 2 (Cognitive Reflection: low vs. high) × 2 (Judgment Type: veracity-judgment vs. sharing-

[3] Exploratory analyses investigating potential differences between self-identified Democrats and self-identified Republicans are presented in the online supplemental materials.

Figure 5

Signal Detection d′ Scores Reflecting Truth Sensitivity in Responses to Political Information as a Function of Ideology-Congruence (Congruent vs. Incongruent), Judgment Type (Veracity Judgment vs. Sharing Decision), and Cognitive Reflection (Low Reflection vs. High Reflection), Experiment 2

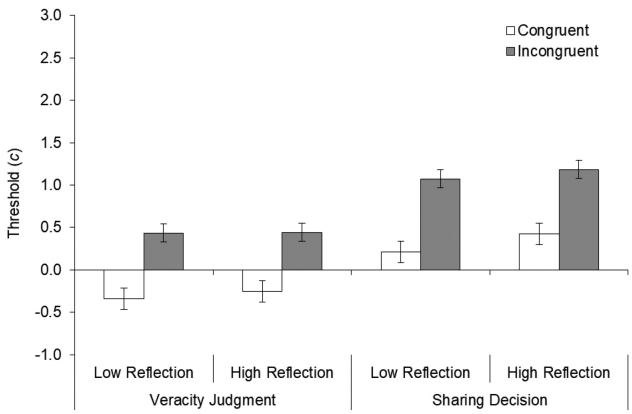


Note. Higher *d′* scores reflect greater truth sensitivity. Error bars depict 95% confidence intervals.

decision) × 2 (Ideology-Congruence: congruent vs. incongruent) ANOVA with the first two variables as between-subjects factors and the third as a within-subject factor. Means and 95% confidence intervals across conditions are presented in Figure 6. Confirming the presence of a partisan-bias effect, the ANOVA revealed a significant main effect of Ideology-Congruence, $F(1, 726) = 543.81$, $p < .001$, $\eta_p^2 = 0.428$, indicating that participants showed a lower acceptance threshold for ideology-congruent headlines compared to ideology-incongruent headlines. Replicating findings of Experiment 1, a significant main effect of Judgment Type further revealed that participants had a higher acceptance threshold for sharing decisions

Figure 6

Signal Detection c Scores Reflecting Response Threshold in Responses to Political Information as a Function of Ideology-Congruence (Congruent vs. Incongruent), Judgment Type (Veracity Judgment vs. Sharing Decision), and Cognitive Reflection (Low Reflection vs. High Reflection), Experiment 2



Note. Higher *c* scores reflect higher acceptance threshold. Error bars depict 95% confidence intervals.

compared to veracity judgments, $F(1, 726) = 179.25$, $p < .001$, $\eta_p^2 = 0.198$. A significant main effect of Cognitive Reflection further revealed that participants in the high-reflection condition had a higher acceptance threshold than participants in the low-reflection condition, $F(1, 726) = 4.60$, $p = .032$, $\eta_p^2 = 0.006$. Yet, consistent with Hypothesis 2, partisan bias was not significantly affected by Cognitive Reflection, as indicated by a non-significant two-way interaction of Cognitive Reflection and Ideology-Congruence, $F(1, 726) = 1.85$, $p = .175$, $\eta_p^2 = 0.003$, and a non-significant three-way interaction of Cognitive Reflection, Ideology-Congruence, and Judgment Type, $F(1, 726) = 0.04$, $p = .845$, $\eta_p^2 < 0.001$. The main effect of Ideology-Congruence was statistically significant for both veracity judgments, $F(1, 359) = 266.44$, $p < .001$, $\eta_p^2 = 0.426$, and sharing decisions, $F(1, 367) = 278.64$, $p < .001$, $\eta_p^2 = 0.432$, indicating a significant partisan-bias effect for both types of judgment. In contrast, the two-way interaction of Cognitive Reflection and Ideology-Congruence was not statistically significant for veracity judgments, $F(1, 359) = 0.73$, $p = .394$, $\eta_p^2 = 0.002$, and sharing decisions, $F(1, 367) = 1.13$, $p = .288$, $\eta_p^2 = 0.003$, indicating that cognitive reflection did not significantly qualify the size of partisan bias in either of the two judgment conditions.

Discussion

In addition to replicating the main findings of Experiment 1, Experiment 2 revealed that truth sensitivity was greater under high-reflection conditions compared to low-reflection conditions (consistent with Hypothesis 1a). However, although this difference emerged for veracity judgments (consistent with Hypothesis 1b), it did not emerge for sharing decisions (inconsistent with Hypothesis 1c). The latter findings may suggest that cognitive reflection influences veracity judgments but not sharing decisions. Yet, the absence of a significant interaction between cognitive reflection and judgment type renders such a conclusion premature. Thus, although the current findings support the idea that cognitive reflection increases truth sensitivity in veracity judgments, they are inconclusive regarding the effect of cognitive reflection on truth sensitivity in sharing decisions. Moreover, the obtained effect of cognitive reflection on truth sensitivity in veracity judgments was relatively weak with an effect size that qualifies as small in terms of current conventions (J. Cohen, 1988). This small effect stands in contrast to the large effects of judgment type on truth sensitivity and response thresholds in Experiments 1 and 2, as well as the large partisan-bias effect in the two studies.

Consistent with Hypothesis 2a, cognitive reflection did not affect partisan bias. This null effect emerged for both veracity judgments and sharing decisions, consistent with Hypotheses 2b and 2c. Thus, although cognitive reflection was effective in increasing truth sensitivity in veracity judgments, it was ineffective in reducing partisan bias. Together, these findings suggest that (a) cognitive reflection can increase the correct identification of true and false information in a proximal way without additional information about the source and (b) enhanced truth sensitivity resulting from cognitive reflection does not lead to reduced levels of partisan bias.

Experiment 3

Experiment 3 aimed to provide deeper insights into the psychological underpinnings of partisan bias in veracity judgments and sharing decisions. A common explanation for partisan bias in

responses to misinformation is that it arises from processes of motivated reasoning (see Kruglanski et al., 2020; Kunda, 1990). According to this view, people accept ideology-congruent information and reject ideology-incongruent information, because doing so supports and protects beliefs that are central to their social identity, which in turn elicits positive feelings about the self (e.g., Van Bavel & Pereira, 2018). Conversely, accepting ideology-incongruent information as true or rejecting ideology-congruent information as false elicits feelings of self-threat, because doing so suggests conclusions that conflict with one's identity-related beliefs. Together with the idea that people want to feel good about themselves (see Alicke & Sedikides, 2009), these assumptions imply that people should be more likely to accept ideology-congruent information than ideology-incongruent information, as found in the current studies.

However, as pointed out by Pennycook and Rand (2021), greater acceptance of ideology-congruent compared to ideology-incongruent information could also be the product of non-motivational, cognitive processes that follow a pattern of Bayesian belief updating (see also Gawronski, 2021; Tappin et al., 2020). According to this view, acceptance of ideology-congruent information and rejection of ideology-incongruent information is not a motivationally driven bias, but a rational decision in response to new information that conflicts with strongly held beliefs (Pennycook & Rand, 2021). To the extent that someone holds strong beliefs about an issue (equivalent to the notion of strong Bayesian priors), it would be rational from a Bayesian view to reconcile such conflicts by searching for an explanation for new belief-incongruent information instead of changing one's prior beliefs. As treating new belief-incongruent information as false resolves the conflict without a need to change one's prior beliefs (see Gawronski, 2012; Johnson-Laird, 2012), these considerations suggest that people should be more likely to accept ideology-congruent information than ideology-incongruent information, as found in the current studies.

The main goal of Experiment 3 was to test the two competing accounts of partisan bias in responses to misinformation. Toward this end, we compared partisan bias in responses to misinformation across conditions in which participants' self was either affirmed or threatened (see G. L. Cohen et al., 2007). Regarding the current question, a valuable aspect of this manipulation is that motivational and cognitive accounts lead to different predictions about its effect on partisan bias in responses to misinformation.

From the perspective of motivational accounts, an important aspect of self-affirmation is that it elicits positive self-feelings, whereas self-threat elicits negative self-feelings (see Sherman & Cohen, 2006). Moreover, positive self-feelings should serve as a buffer against potential threats, which should increase people's openness to potentially threating information (e.g., Sherman et al., 2000). Conversely, negative self-feelings should make people especially sensitive to potential threats, which should reduce their openness to potentially threatening information (e.g., Liberman & Chaiken, 1992). Thus, combined with the ideas that ideology-incongruent information poses a threat to one's social identity and ideology-congruent information supports one's social identity, partisan bias in responses to misinformation should be more pronounced under conditions of self-threat compared to conditions of self-affirmation.

From the perspective of cognitive accounts, an important aspect of self-affirmation is that it increases self-confidence, whereas self-threat reduces self-confidence (see Briñol & Petty, 2022). Moreover, whereas higher confidence in one's personally held beliefs should increase the likelihood that new belief-incongruent information is dismissed as false, lower confidence should increase the likelihood that people question their personally held beliefs in response to new belief-incongruent information. Thus, combined with the fact that ideology-incongruent information is, by definition, more likely to conflict with prior beliefs than ideology-congruent information, partisan bias in responses to misinformation should be more pronounced under conditions of self-affirmation compared to conditions of self-threat.

Based on the additional assumption that effects of self-affirmation versus self-threat should be limited to partisan bias without affecting truth sensitivity, these considerations led to the following preregistered hypotheses:

> *Hypothesis 1:* Self-affirmation (vs. self-threat) will have no effect on truth sensitivity captured by SDT's d' index (Hypothesis 1a), and this null effect will be observed for both veracity judgments (Hypothesis 1b) and sharing decisions (Hypothesis 1c).
>
> *Hypothesis 2:* In line with predictions derived from motivational accounts of partisan bias, self-affirmation (vs. self-threat) will decrease partisan bias captured by the difference between SDT's *c* index for ideology-congruent and ideology-incongruent news headlines (Hypothesis 2a), and this decrease will be observed for both veracity judgments (Hypothesis 2b) and sharing decisions (Hypothesis 2c).
>
> *Hypothesis 3:* In line with predictions derived from cognitive accounts of partisan bias, self-affirmation (vs. self-threat) will increase partisan bias captured by the difference between SDT's *c* index for ideology-congruent and ideology-incongruent news headlines (Hypothesis 3a), and this increase will be observed for both veracity judgments (Hypothesis 3b) and sharing decisions (Hypothesis 3c).

Although our preregistered hypotheses were formulated in an either-or fashion, it is worth nothing that the mechanisms proposed by motivational and cognitive accounts are not mutually exclusive. It is entirely possible that the two mechanisms operate at the same time, which should lead to an overall null effect on partisan bias because the outcomes of the two mechanisms would compensate each other. Whereas enhanced positive self-feelings resulting from self-affirmation (vs. self-threat) would increase partisan bias, greater self-confidence resulting from self-affirmation (vs. self-threat) would decrease partisan bias, leading to an overall null effect of our experimental manipulation. To explore the possibility of such compensatory effects, we preregistered additional exploratory analyses using measures of self-feelings and self-confidence as simultaneous mediators in a multiple-regression mediator analysis. Toward this end, we preregistered that we would (a) first test whether self-affirmation (vs. self-threat) significantly increases positive self-feelings (related to the prediction of motivational accounts) and whether self-affirmation (vs. self-threat) significantly increases self-confidence (related to the prediction of cognitive accounts), and then (b) simultaneously regress partisan bias onto dummy-coded conditions of self-affirmation versus self-threat, standardized scores of self-feelings, and standardized scores of self-confidence. According to motivational accounts, self-affirmation (vs. self-threat) should increase positive self-feelings, and positive self-feelings should show a negative relationship with partisan bias. According

to cognitive accounts, self-affirmation (vs. self-threat) should increase self-confidence, and greater self-confidence should show a positive relationship with partisan bias. In line with the above reasoning about multi-faceted effects of our experimental manipulation, these results may emerge even when the experimental manipulation itself has no significant effect on partisan bias.

To address these questions, self-identified Republicans and self-identified Democrats were presented with true and false news headlines that had either a pro-Republican or pro-Democrat slant. Following the procedure in Experiment 1, half of the participants were asked to indicate for each headline if it is true or false; the remaining half were asked to indicate for each headline whether they would share the story online. Orthogonal to the manipulation of judgment type, half of the participants were asked to explain prior to the headline-judgment task why an important personal value is meaningful to them (self-affirmation condition); the remaining half were asked to describe a time when they failed to live up to an important personal value (self-threat condition). Responses were analyzed using SDT to quantify the degree of truth sensitivity and partisan bias in veracity judgments and sharing decisions, respectively.

Method

Preregistration

The design, hypotheses, and analysis plan of Experiment 3 were pre-registered prior to data collection at https://osf.io/fpqrj/. The data for Experiment 3 were collected in June 2022.

Participants and Design

We aimed to have 95% power for the detection of a small effect of $f = 0.10$ in a 2 (Self: self-affirmation vs. self-threat) × 2 (Judgment Type: veracity-judgment vs. sharing-decision) × 2 (Ideology-Congruence: congruent vs. incongruent) mixed ANOVA with the first two variables as between-subjects factors and the last one as a within-subject factor (two-tailed). Assuming a correlation between measures of $r = .30$ and using a nonsphericity correction of $\varepsilon = 1$, a sample size of at least 608 participants is required to meet this target. Anticipating that approximately 10% of the participants may fail to pass our attention check and to obtain sufficient power for slightly smaller effects, we set our preregistered target sample to 800 participants prior to exclusions. Participants were recruited on Prolific Academic via two assignments for the separate recruitment of 400 self-identified Democrats and 400 self-identified Republicans. Eligibility criteria for participation were identical to Experiments 1 and 2. The study took approximately 15–20 min to complete, and participants were compensated $4.00 for their time.

Following our preregistered stopping rule, data collection ended once 800 participants had been approved for credit. Of the 869 Prolific workers who started the study, 801 completed all measures.[4] Of the 801 participants with complete data, 39 failed to pass an attention check and 11 reported inconsistent political affiliations in Prolific's prescreening survey and the current study. Data from these participants were excluded from analyses. Thus, the final sample included a total of 751 participants (292 men, 453 women, two prefer not to answer, four other), 384 of which identified as Democrat and 367 of which identified as Republican ($n = 189$ in the veracity-judgment/self-affirmation condition; $n = 192$ in the veracity-judgment/self-threat condition; $n = 188$ in the sharing-decision/self-affirmation condition; $n = 182$ in the sharing-decision/self-threat condition).[5] Participants' age ranged from 18 to 79 years ($M_{age} = 37.25$ years, $SD_{age} = 13.48$). Of the 751 participants in the final sample, 625 identified as White, 46 identified as Black or African American, zero identified as American Indian or Alaska Native, 40 identified as Asian, zero identified as Native Hawaiian or Pacific Islander, 14 as other, and 26 identified with more than one race category. For household income, 80 reported incomes lower than $20,000, 140 reported incomes between $20,000 and $40,000, 137 reported incomes between $40,000 and $60,000, 118 reported incomes between $60,000 and $80,000, 104 reported incomes between $80,000 and $100,000, and 172 reported incomes higher than $100,000. For education, two reported having less than a high school degree, 86 reported having a high school degree or equivalent, 172 reported having some college education with no degree, 78 reported having a two-year college associate degree, 301 reported having a 4-year college bachelor's degree, 90 reported having a master's degree, six reported having a doctoral degree, and 16 reported having a professional JD or MD degree. Self-identified Republicans considered themselves significantly more conservative than self-identified Democrats in general (Ms = 5.66 vs. 1.85), $t(729.53) = 54.66$, $p < .001$, $d = 3.99$, in terms of economic issues (Ms = 5.75 vs. 2.08), $t(749) = 44.79$, $p < .001$, $d = 3.27$, and in terms of social issues (Ms = 5.22 vs. 1.70), $t(639.87) = 38.88$, $p < .001$, $d = 2.84$.

The study included a 2 (Headline Accuracy: true vs. false) × 2 (Headline Slant: pro-Democrat vs. pro-Republican) × 2 (Political Affiliation: Democrat vs. Republican) × 2 (Judgment Type: veracity-judgment vs. sharing-decision) × 2 (Self: self-affirmation vs. self-threat) mixed design with the first two factors varying within-subjects and the latter three factors varying between-subjects. Participants were randomly assigned to one of the four between-subjects conditions.

Procedure, Materials, and Measures

The procedures, materials, and measures were identical to Experiment 2 with three exceptions. First, we replaced the manipulation of cognitive reflection with a manipulation of self-affirmation versus self-threat (adapted from G. L. Cohen et al., 2007). Second, we included two items to test the effectiveness of our self-affirmation versus self-threat manipulation in influencing self-feelings and self-confidence. Third, we used an updated set of headlines in which outdated headlines were replaced with new headlines that underwent the same screening procedures utilized for the selection of headlines in Experiments 1 and 2 (see Appendix). The results of the second pilot study to identify suitable headlines and the list of headlines used in the current study can be found at https://osf.io/d2rne/.

Participants in the self-affirmation condition were asked to rank 11 characteristics and values via drag-and-drop in order of their

[4] One participant's submission was initially rejected because the participant submitted an incorrect completion code. The participant was granted compensation retroactively after providing evidence for their participation by answering several questions about the study contents. The data from this participant were retained in the final sample.

[5] Exploratory analyses investigating potential differences between self-identified Democrats and self-identified Republicans are presented in the online supplemental materials.

importance to the participant from 1 to 11 (1 = *most important item*, 11 = *least important item*). The list of items included: *artistic skills/aesthetic appreciation*, *sense of humor*, *relationships with friends/family*, *spontaneity/living life in the moment*, *social skills*, *athletics*, *musical ability/appreciation*, *physical attractiveness*, *creativity*, *business/managerial skills*, and *romantic values*. After participants completed the ranking task, they were presented with their highest-ranked value and asked to explain why it is meaningful to them. Participants in the self-threat condition completed the same value-ranking task but were instead asked to describe a time when they failed to live up to their highest-ranked value. Participants in both conditions had 5 min to complete the writing task.

After completing the writing task to which participants had been randomly assigned, participants responded to two questions to gauge the effectiveness of our self-affirmation versus self-threat manipulation in influencing self-feelings and self-confidence. The first question was *How do you feel about yourself?* which participants answered on a 5-point rating scale ranging from 1 (*very negative*) to 5 (*very positive*). The second question was *How confident do you feel about your personal views?* which participants answered on a 5-point rating scale ranging from 1 (*not at all confident*) to 5 (*extremely confident*).[6] The order of the two measures was counterbalanced across participants.

Data Aggregation and Treatment

The preregistered data aggregation plan followed the procedures of Experiment 2; the preregistered exclusion criteria were the same as in Experiment 1.

Results

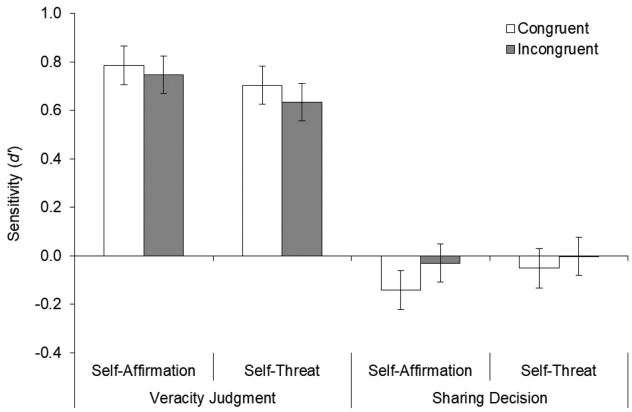
Confirmatory Analyses

Truth Sensitivity. To investigate effects of self-affirmation versus self-threat on truth sensitivity as a function of veracity judgments versus sharing decisions, d' scores were submitted to a 2 (Self: self-affirmation vs. self-threat) × 2 (Judgment Type: veracity-judgment vs. sharing-decision) × 2 (Ideology-Congruence: congruent vs. incongruent) ANOVA with the first two variables as between-subjects factors and the third as a within-subject factor (see Figure 7). Replicating results of Experiments 1 and 2, a significant main effect of Judgment Type indicated that truth sensitivity was greater for veracity judgments than sharing decisions, $F(1, 747) = 575.81$, $p < .001$, $\eta_p^2 = 0.435$. This main effect was qualified by a significant two-way interaction between Judgment Type and Ideology-Congruence, $F(1, 747) = 7.79$, $p = .005$, $\eta_p^2 = 0.010$, and a significant two-way interaction between Judgment Type and Self, $F(1, 747) = 5.88$, $p = .016$, $\eta_p^2 = 0.008$. Further analyses revealed that the observed difference between veracity judgments and sharing decisions was more pronounced for ideology-congruent headlines, $F(1, 747) = 427.36$, $p < .001$, $\eta_p^2 = 0.364$, compared to ideology-incongruent headlines, $F(1, 747) = 316.49$, $p < .001$, $\eta_p^2 = 0.298$, and under self-affirmation conditions, $F(1, 375) = 365.10$, $p < .001$, $\eta_p^2 = 0.493$, compared to self-threat conditions, $F(1, 372) = 222.71$, $p < .001$, $\eta_p^2 = 0.374$. No other main or interaction effect reached statistical significance (all Fs < 1, all ps > .32).

Response Threshold. To investigate effects of self-affirmation versus self-threat on partisan bias as a function of veracity judgments versus sharing decisions, c scores were submitted to a 2 (Self: self-affirmation vs. self-threat) × 2 (Judgment Type: veracity-judgment vs. sharing-decision) × 2 (Ideology-Congruence: congruent vs. incongruent) ANOVA with the first two variables as between-subjects factors and the third variable as a within-subject factor (see Figure 8). Replicating findings of Experiments 1 and 2, a significant main effect of Judgment Type revealed that participants had a higher acceptance threshold for sharing decisions compared to veracity judgments, $F(1, 747) = 294.74$, $p < .001$, $\eta_p^2 = 0.283$. Moreover, confirming the presence of a partisan-bias effect, the ANOVA revealed a significant main effect of Ideology-Congruence, $F(1, 747) = 522.73$, $p < .001$, $\eta_p^2 = 0.412$, indicating that participants showed a lower acceptance threshold for ideology-congruent headlines compared to ideology-incongruent headlines. These main effects were qualified by a significant two-way interaction between Ideology-Congruence and Judgment Type, $F(1, 747) = 7.20$, $p = .007$, $\eta_p^2 = 0.010$, indicating that partisan bias was more pronounced for veracity judgments, $F(1, 379) = 375.18$, $p < .001$, $\eta_p^2 = 0.497$, than sharing decisions, $F(1, 368) = 179.04$, $p < .001$, $\eta_p^2 = 0.327$. Counter to Hypotheses 2 and 3, the main effect of Ideology-Congruence was not qualified by a significant two-way interaction with the manipulation of self-affirmation versus

Figure 7

Signal Detection d' Scores Reflecting Truth Sensitivity in Responses to Political Information as a Function of Ideology-Congruence (Congruent vs. Incongruent), Judgment Type (Veracity Judgment vs. Sharing Decision), and Self (Self-Affirmation vs. Self-Threat), Experiment 3

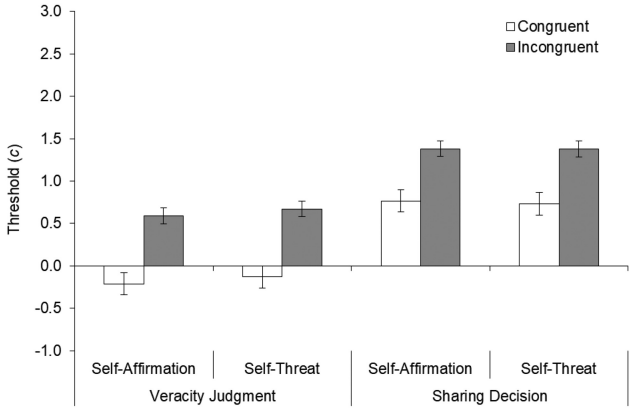


Note. Higher d' scores reflect greater truth sensitivity. Error bars depict 95% confidence intervals.

[6] Prior to collecting the data for Experiment 3, we conducted two similar studies that used a control condition with a neutral task instead of a task designed to induce self-threat (see McQueen & Klein, 2006). The first study did not produce any significant effects on the two measures of self-feelings and self-confidence; the second study revealed no significant effect on the measure of self-feelings and a significant effect on the measure of self-confidence with an effect size that fell below the conventional benchmark for a small effect ($d < .20$; see J. Cohen, 1988). Because these findings suggest that the manipulation in the two studies was largely ineffective in inducing the intended differences in self-feelings and self-confidence, we conducted the current Experiment 3 with a self-threat induction in the control condition instead of a neutral control task. The preregistrations, data, analysis files, and a summary of the results of the two additional studies are available at https://osf.io/43hsg/.

Figure 8

Signal Detection c *Scores Reflecting Response Threshold in Responses to Political Information as a Function of Ideology-Congruence (Congruent vs. Incongruent), Judgment Type (Veracity Judgment vs. Sharing Decision), and Self (Self-Affirmation vs. Self-Threat), Experiment 3*



Note. Higher *c* scores reflect higher acceptance threshold. Error bars depict 95% confidence intervals.

self-threat, $F(1, 747) = 0.05$, $p = .824$, $\eta_p^2 < 0.001$. There was also no significant three-way interaction between Ideology-Congruence, Judgment Type, and Self, $F(1, 747) = 0.09$, $p = .770$, $\eta_p^2 < 0.001$. The main effect of Ideology-Congruence was statistically significant for both veracity judgments and sharing decisions (see above), and not qualified by self-affirmation versus self-threat for either judgment type (all Fs < 1, all ps $> .73$).

Exploratory Analyses

Self-Feelings and Self-Confidence. The manipulation of self-affirmation versus self-threat was effective in influencing both self-confidence and self-feelings. First, self-confidence was significantly greater in the self-affirmation condition than in the self-threat condition (Ms $= 4.17$ vs. 3.88, respectively), $t(716.03) = 4.80$, $p < .001$, $d = 0.351$. Second, self-feelings were significantly more positive in the self-affirmation condition than in the self-threat condition (Ms $=$ 3.75 vs. 3.34, respectively), $t(732.82) = 6.05$, $p < .001$, $d = 0.442$. Self-confidence and positive self-feelings were positively correlated across conditions ($r = .516$, $p < .001$). The effects of self-affirmation versus self-threat on self-confidence and self-feelings remained statistically significant when controlling for the correlation between the two measured variables, $F(1, 748) = 4.09$, $p = .044$, $\eta_p^2 = 0.005$ for self-confidence and $F(1, 748) = 17.23$, $p < .001$, $\eta_p^2 = 0.023$ for self-feelings.

Prediction of Partisan Bias. Because our manipulation of self-affirmation versus self-threat was effective in influencing both self-confidence and self-feelings, we conducted preregistered exploratory analyses to investigate the possibility that the null effect of self-affirmation versus self-threat on partisan bias concealed compensatory influences of cognitive and motivational processes. Whereas cognitive accounts suggest that enhanced self-confidence resulting from self-affirmation (vs. self-threat) should increase partisan bias, motivational accounts suggest that enhanced positive self-feelings resulting from self-affirmation (vs. self-threat) should decrease partisan bias, leading to an overall null effect of self-affirmation (vs. self-threat) on partisan bias. To test this idea, we simultaneously regressed partisan bias onto dummy-coded conditions of self-affirmation versus self-threat, standardized scores of self-feelings, and standardized scores of self-confidence, following our preregistered plan for exploratory analyses. Analyses were conducted separately for veracity judgments and sharing decisions. For veracity judgments, self-confidence showed a significant positive association with partisan bias ($\beta = 0.168$, $p = .005$) and positive self-feelings showed a marginal negative association with partisan bias ($\beta = -0.111$, $p = .062$). For sharing decisions, partisan bias was not significantly related to self-confidence ($\beta = 0.084$, $p = .174$) and self-feelings ($\beta = 0.050$, $p = .427$).[7]

Expanding on the results of the multiple-regression analyses, we also conducted exploratory path-model analyses for the full multiple-mediation model using structural equation modeling (SEM) in *MPlus* Version 8.0 (Muthén & Muthén, 2017). Figure 9 depicts the results of the path-model analyses for veracity judgments (upper panel) and sharing decisions (lower panel).

For veracity judgments, the manipulation of self-affirmation versus self-threat showed a significant effect on both self-confidence and self-feelings. Moreover, whereas self-confidence showed a significant positive association with partisan bias, positive self-feelings showed a marginal negative association with partisan bias. The indirect path of self-affirmation versus self-threat on partisan bias was statistically significant for the mediation via self-confidence ($Z = 2.32$, $p = .020$) and marginal for the mediation via self-feelings ($Z = 1.72$, $p = .086$). These results support the assumptions of cognitive accounts, but they are inconclusive regarding the assumptions of motivational accounts. Based on the marginal association between self-feelings and partisan bias and the marginal indirect effect, we can neither accept nor reject the idea that self-affirmation versus self-threat reduced partisan bias by enhancing positive self-feelings.

For sharing decisions, the manipulation of self-affirmation versus self-threat showed a significant effect on both self-confidence and self-feelings. Different from the results for veracity judgments, partisan bias was not significantly related to self-confidence and self-feelings. The indirect effect of self-affirmation versus self-threat on partisan bias was not statistically significant in either case (all Zs < 1.23, all ps $> .291$).

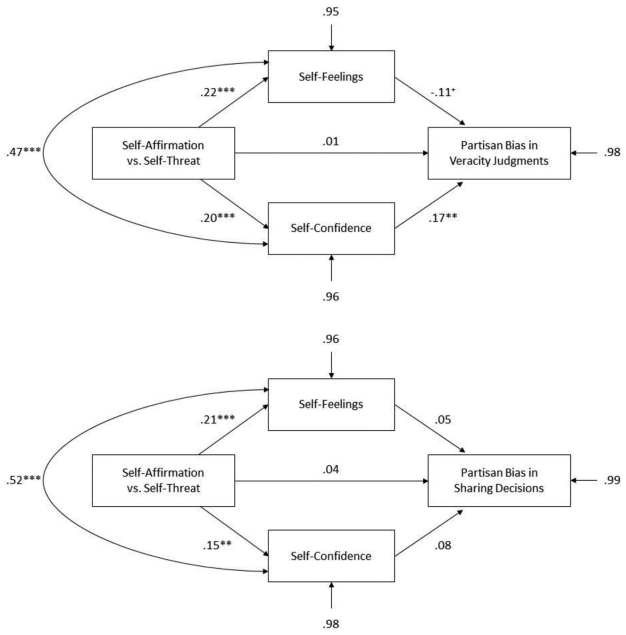
Discussion

In addition to replicating the main findings of Experiment 1 (see also Experiment 2), Experiment 3 revealed three sets of noteworthy findings. First, counter to Hypotheses 2 and 3, our manipulation of self-affirmation versus self-threat had no significant effect on partisan bias. According to motivational accounts, partisan bias in responses to misinformation should be more pronounced under conditions of self-threat compared to conditions of self-affirmation. In

[7] To investigate if the obtained results are unique to partisan bias, we also conducted corresponding analyses with overall truth sensitivity as the criterion, calculated as the average of truth sensitivity for ideology-congruent and ideology-incongruent information (see Experiment 1). Neither self-confidence nor self-feelings showed a significant association with overall truth sensitivity (all $|\beta$s$| < 0.060$, all ps $> .158$).

Figure 9

Results of the Path-Model Analyses for Effects Self-Affirmation (vs. Self-Threat) on Partisan Bias in Veracity Judgments (Upper Panel) and Sharing Decisions (Lower Panel) via Positive Self-Feelings and Levels of Self-Confidence, Experiment 3



contrast, according to cognitive accounts, partisan bias in responses to misinformation should be more pronounced under conditions of self-affirmation compared to conditions of self-threat. Although measures of self-feelings and self-confidence provided strong support for the effectiveness of our experimental manipulation of self-affirmation versus self-threat, neither of the two competing hypotheses received empirical support in our preregistered confirmatory analyses. There was also no significant effect of self-affirmation versus self-threat on truth sensitivity (consistent with Hypothesis 1).

Second, preregistered exploratory analyses provided support for a cognitive explanation of partisan bias. According to cognitive accounts, partisan bias should increase as a function of self-confidence, which is consistent with the findings that (a) self-confidence was positively associated with partisan bias and (b) self-affirmation versus self-threat showed an indirect positive effect on partisan bias via enhanced self-confidence. However, this association emerged only for veracity judgments and did not generalize to sharing decisions. Although the exploratory nature of these analyses and the lack of generality across judgment types suggest caution in interpreting this finding, it is worth noting that exploratory analyses in Experiment 1 revealed a similar pattern using a measure of self-perceived ability in identifying news that is made up. Specifically, we found that self-perceived ability in identifying made-up news was positively associated with partisan bias in veracity judgments, but not sharing decisions. To the extent that the self-perceived ability in identifying made-up news can be interpreted as an indicator of confidence (see Lyons et al., 2021), the convergence of the two findings provides preliminary evidence for the idea that partisan bias in veracity judgments might be the product of basic cognitive processes that follow a pattern of Bayesian belief updating (see Pennycook & Rand, 2021; Tappin et al., 2020). However, it remains unclear what underlies partisan bias in sharing decisions, which might be driven by mechanisms that are different from the ones underlying partisan bias in veracity judgments.

Third, the results of Experiment 3 provide no evidence for the hypothesis that partisan bias in sharing decisions decreases as a function of positive self-feelings, as suggested by motivational accounts which assume that partisan bias arises from a desire to support and protect beliefs that are central to one's social identity (see Van Bavel & Pereira, 2018). Regarding partisan bias in veracity judgments, the current findings are inconclusive about a potential role of self-feelings, given that the negative association between positive feelings and partisan bias as well as the indirect effect of self-affirmation versus self-threat via self-feelings were only marginal. These findings do not provide a sufficiently strong basis to either accept or reject the idea that self-affirmation versus self-threat reduced partisan bias by enhancing positive self-feelings.

Experiment 4

The three preceding experiments consistently revealed that, although partisan bias in responses to misinformation was similarly pronounced for veracity judgments and sharing decisions, truth sensitivity was substantially greater for veracity judgments compared to sharing decisions. Moreover, whereas truth sensitivity in veracity judgments was greater than chance in every single case, there was not a single case where truth sensitivity in sharing decisions was greater than chance (see Figures 3, 5, and 7). Together, these results suggest that, although participants were able to accurately distinguish between true and false information to a considerable extent, information veracity had no impact whatsoever on sharing decisions. Our finding that partisan bias does not differ for veracity judgments and sharing decisions further suggests that the observed discrepancy in truth sensitivity is not driven by greater partisan bias in information sharing. Instead, it seems more likely that the low degree of truth sensitivity in sharing decisions results from a lack of attention to information veracity (see Pennycook & Rand, 2021). These considerations raise the question of whether the quality of shared information could be increased by directing attention to information veracity prior to a potential sharing decision. Consistent with this idea, several studies have found that the quality of shared information increased when participants had been nudged to think about the veracity of the focal information (e.g., Pennycook, Epstein, et al., 2021; Pennycook et al., 2020). However, the reliability of such truth-prompt effects has been questioned by failed replications (Roozenbeek et al., 2021) and reanalyses of existing data suggesting that effects of truth prompts are limited to participants with liberal political ideology and do not generalize to participants with conservative political ideology (Rathje et al., 2022; but see Pennycook & Rand, 2022).

The main goal of Experiment 4 was to gain deeper insights into the potential effects of truth prompts by using an SDT approach to distinguish between truth sensitivity and partisan bias in sharing decisions. Based on our finding that veracity judgments and sharing decisions differ in terms of truth sensitivity, but not partisan bias, we aimed to test the hypotheses that truth prompts may be effective in increasing truth sensitivity but ineffective in reducing partisan bias. The two hypotheses were formally preregistered as follows:

> *Hypothesis 1:* Truth prompts will increase truth sensitivity in sharing decisions as captured by SDT's d' index.

Hypothesis 2: Truth prompts will have no effect on partisan bias in sharing decisions as captured by the difference between SDT's *c* index for ideology-congruent and ideology-incongruent news headlines.

To test these hypotheses, self-identified Republicans and self-identified Democrats were presented with true and false news headlines that had either a pro-Republican or a pro-Democrat slant. Participants were asked to indicate for each headline whether they would share the story online. To investigate the impact of truth prompts on truth sensitivity and partisan bias in sharing decisions, half of the participants were asked to judge each headline's veracity prior to answering the question about sharing. The remaining half answered the sharing question without being asked about the headlines' veracity. Responses were analyzed using SDT to quantify the degree of truth sensitivity and partisan bias in sharing decisions.

Method

Preregistration

The design, hypotheses, and analysis plan of Experiment 4 were preregistered prior to data collection at https://osf.io/wme9r/. The data for Experiment 4 were collected in July 2021.

Participants and Design

We aimed to have at least 95% power for the detection of a small between-group difference of $d = 0.30$ in a *t*-test for independent means (two-tailed), which requires a sample of 580 participants. For the critical tests in the current study, a sample of this size provides a power of 95% in detecting a small effect of $f = 0.087$ in a 2×2 mixed ANOVA with one factor varying between-subjects and the other varying within-subjects (two-tailed), assuming a correlation between measures of $r = .30$ and using a nonsphericity correction of $\varepsilon = 1$. Anticipating that approximately 10% of the participants may fail to pass our attention check, we set our preregistered target sample to 640 participants prior to exclusions. Participants were recruited on Prolific Academic via two assignments for the separate recruitment of 320 self-identified Democrats and 320 self-identified Republicans. Eligibility criteria for participation were identical to Experiments 1, 2, and 3. The study took approximately 10–15 min to complete, and participants were compensated $3.00 for their time.

Following our preregistered stopping rule, data collection ended once 640 participants had been approved for credit. Of the 652 Prolific workers who started the study, 641 completed all measures.[8] Of the 641 participants with complete data, 65 failed to pass an attention check and 7 reported inconsistent political affiliations in Prolific's prescreening survey and the current study. Data from these participants were excluded from analyses. Thus, the final sample included a total of 569 participants (261 men, 298 women, three prefer not to answer, and seven other), 279 of which identified as Democrat and 290 of which identified as Republican ($n = 280$ in the truth-prompt-absent condition; $n = 289$ in the truth-prompt-present condition).[9] Participants' age ranged from 18 to 74 years ($M_{age} = 35.94$ years, $SD_{age} = 12.67$). Of the 569 participants in the final sample, 448 identified as White, 54 identified as Black or African American, one identified as American Indian or Alaska Native, 34 identified as Asian, zero identified as Native Hawaiian or Pacific Islander, 14 as other, and 18 identified with more than one race category. For household income, 48 reported incomes lower than $20,000, 117 reported incomes between $20,000 and $40,000, 99 reported incomes between $40,000 and $60,000, 101 reported incomes between $60,000 and $80,000, 57 reported incomes between $80,000 and $100,000, and 147 reported incomes higher than $100,000. For education, 1 reported having less than a high school degree, 73 reported having a high school degree or equivalent, 120 reported having some college education with no degree, 51 reported having a 2-year college associate degree, 199 reported having a 4-year college bachelor's degree, 101 reported having a master's degree, 11 reported having a doctoral degree, and 13 reported having a professional JD or MD degree. Self-identified Republicans considered themselves significantly more conservative than self-identified Democrats in general (Ms = 5.58 vs. 2.02), $t(567) = 36.25$, $p < .001$, $d = 3.04$, in terms of economic issues (Ms = 5.72 vs. 2.22), $t(553.95) = 31.96$, $p < .001$, $d = 2.68$, and in terms of social issues (Ms = 5.21 vs. 1.84), $t(547.35) = 30.21$, $p < .001$, $d = 2.53$.

The study included a 2 (Headline Accuracy: true vs. false) × 2 (Headline Slant: pro-Democrat vs. pro-Republican) × 2 (Political Affiliation: Democrat vs. Republican) × 2 (Truth Prompt: present vs. absent) mixed design with the first two factors varying within-subjects and the latter two factors varying between-subjects. Participants were randomly assigned to one of the two between-subjects conditions.

Procedure, Materials, and Measures

The procedures, materials, and measures were identical to the sharing-decision condition in Experiment 1 with two exceptions. First, Experiment 4 included a manipulation of truth nudging in that half of the participants were asked to judge the accuracy of each headline before they were asked if they would share the headline online (truth-prompt-present condition). The remaining half were asked if they would share the headlines on social media without being initially asked to judge their accuracy. The accuracy question was identical to the one in the veracity-judgment condition in Experiment 1. Second, instead of randomizing the order of the headlines individually for each participant, the headlines in Experiment 4 were presented in a fixed random order that was held constant for all participants. This change was implemented for technical reasons to permit a direct matching of veracity judgments and sharing decisions in the data file.

Data Aggregation and Treatment

The preregistered data aggregation plan and exclusion criteria were identical to Experiment 3.

[8] One participant's submission was initially rejected, because the participant submitted an incorrect completion code. The participant was granted compensation retroactively after providing evidence for their participation by answering several questions about the study contents. The data from this participant were retained in the final sample.

[9] Exploratory analyses investigating potential differences between self-identified Democrats and self-identified Republicans are presented in the online supplemental materials.

Results

Confirmatory Analyses

Truth Sensitivity. To investigate effects of truth nudging on truth sensitivity in sharing decisions, d' scores were submitted to a 2 (Truth Prompt: present vs. absent) × 2 (Ideology-Congruence: congruent vs. incongruent) ANOVA with the first variable as a between-subjects factor and the second as a within-subject factor (see Figure 10). Consistent with Hypothesis 1, a significant main effect of Truth Prompt indicated that truth sensitivity was greater in the truth-prompt-present condition than in the truth-prompt-absent condition, $F(1, 567) = 5.70$, $p = .017$, $\eta_p^2 = 0.010$. There was also a significant main effect of Ideology-Congruence, $F(1, 567) = 10.71$, $p = .001$, $\eta_p^2 = 0.019$, indicating that truth sensitivity was greater for ideology-congruent than ideology-incongruent headlines.

Response Threshold. To investigate effects of truth nudging on partisan bias in sharing decisions, c scores were submitted to a 2 (Truth Prompt: present vs. absent) × 2 (Ideology-Congruence: congruent vs. incongruent) ANOVA with the first variable as a between-subjects factor and the second as a within-subject factor (see Figure 11). Confirming the presence of a partisan-bias effect, the ANOVA revealed a significant main effect of Ideology-Congruence, $F(1, 567) = 369.76$, $p < .001$, $\eta_p^2 = 0.395$, indicating that participants showed a lower threshold for sharing ideology-congruent headlines compared to ideology-incongruent headlines. A significant main effect of Truth Prompt further revealed that participants had a higher acceptance threshold for sharing headlines in the truth-prompt-present condition than in the truth-prompt-absent condition, $F(1, 567) = 59.15$, $p < .001$, $\eta_p^2 = 0.094$. Counter to Hypothesis 2, these main effects were qualified by significant two-way interaction of Ideology-Congruence and Truth Prompt, $F(1, 567) = 32.05$, $p < .001$, $\eta_p^2 = 0.054$, indicating that partisan bias was smaller in the truth-prompt-present condition, $F(1, 288) = 135.88$, $p < .001$, $\eta_p^2 = 0.321$, compared to the truth-prompt absent condition, $F(1, 279) = 230.01$, $p < .001$, $\eta_p^2 = 0.453$.

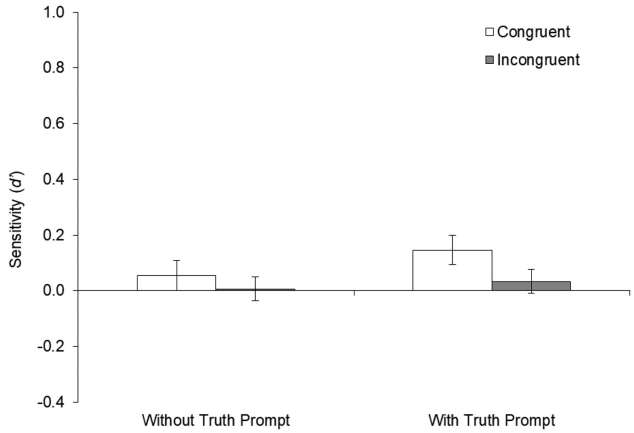
Exploratory Analyses

In addition to the preregistered confirmatory analyses, we conducted two sets of exploratory analyses. First, we conducted exploratory analyses comparing veracity judgments and sharing decisions within the truth-prompt present condition, which provides information on the extent to which sharing decisions do or do not align with prior veracity judgments. Second, we conducted exploratory analyses comparing veracity judgments in the truth-prompt present condition to sharing decisions in the truth-prompt-absent condition, corresponding to the main analyses in Experiment 1. Both analyses were conducted for truth sensitivity and response threshold.

Truth Sensitivity. To compare the levels of truth sensitivity in veracity judgments and sharing decisions among participants who made veracity judgments prior to sharing decisions, we submitted d' scores of participants in the truth-prompt-present condition to a 2 (Judgment Type: veracity-judgment vs. sharing-decision) × 2 (Ideology-Congruence: congruent vs. incongruent) ANOVA with both factors varying within-subjects. Means and 95% confidence intervals in the four conditions are presented in Table 2. The ANOVA revealed a significant main effect of Judgment Type, $F(1, 288) = 275.39$, $p < .001$, $\eta_p^2 = 0.489$, indicating that truth sensitivity was smaller for sharing decisions than veracity judgments even when participants made veracity judgments immediately before they made a sharing decision. There was also a significant main effect of Ideology-Congruence, $F(1, 288) = 6.74$, $p = .010$, $\eta_p^2 = 0.023$, indicating that truth sensitivity was greater for ideology-congruent headlines than ideology-incongruent headlines. The two-way interaction between Judgment Type and Ideology-Congruence was not statistically significant, $F(1, 288) = 2.38$, $p = .124$, $\eta_p^2 = 0.008$.

Figure 10

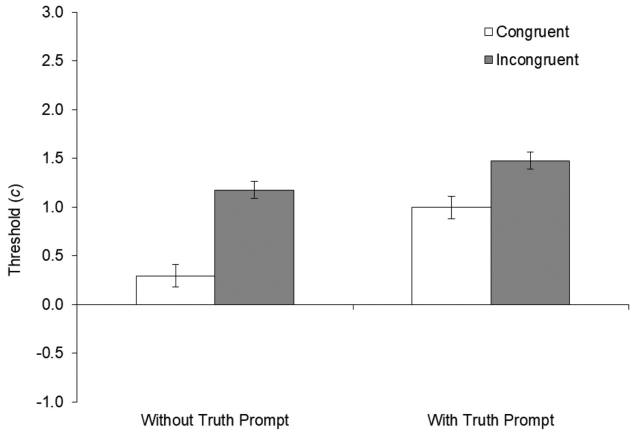
Signal Detection d′ *Scores Reflecting Truth Sensitivity in Sharing Political Information as a Function of Ideology-Congruence (Congruent vs. Incongruent) and Truth Prompt (Without Truth Prompt vs. with Truth Prompt), Experiment 4*



Note. Higher d' scores reflect greater truth sensitivity. Error bars depict 95% confidence intervals.

Figure 11

Signal Detection c *Scores Reflecting Response Threshold in Sharing Political Information as a Function of Ideology-Congruence (Congruent vs. Incongruent) and Truth Prompt (Without Truth Prompt vs. With Truth Prompt), Experiment 4*



Note. Higher c scores reflect higher acceptance threshold. Error bars depict 95% confidence intervals.

Table 2
Means and 95% Confidence Intervals of d′ *and* c *Scores for Veracity Judgments and Sharing Decisions as a Function of Ideology-Congruence (Ideology-Congruent vs. Ideology-Incongruent)*

	Veracity judgments prior to sharing decisions		Sharing decisions with prior veracity judgments	
SDT score	*M*	95% CI	*M*	95% CI
d′				
Ideology-congruent	0.580	[0.509, 0.652]	0.146	[0.093, 0.199]
Ideology-incongruent	0.545	[0.471, 0.619]	0.034	[0.000, 0.067]
C				
Ideology-congruent	−0.234	[−0.308, −0.161]	0.995	[0.893, 1.098]
Ideology-incongruent	0.567	[0.491, 0.642]	1.475	[1.397, 1.553]

Note. Within-participant comparison of scores for participants who provided veracity judgments prior to sharing decisions, Experiment 4; SDT = signal detection theory.

Expanding on these analyses, we also submitted d' scores for veracity judgments in the truth-prompt-present condition and d' scores for sharing decisions in the truth-prompt-absent condition a 2 (Judgment Type: veracity-judgment vs. sharing-decision) × 2 (Ideology-Congruence: congruent vs. incongruent) mixed ANOVA with the first factor varying between-subjects and the second varying within-subjects (corresponding to the main analysis in Experiment 1). Means and 95% confidence intervals in the four conditions are presented in Table 3. Replicating the pattern obtained in the previous studies, the ANOVA revealed a significant main effect of Judgment Type, $F(1, 567) = 219.46$, $p < .001$, $\eta_p^2 = 0.279$, indicating that truth sensitivity was greater for veracity judgments than sharing decisions. No other main or interaction effect reached statistical significance (all Fs < 2.17, all ps $> .14$)

Response Threshold. To compare response thresholds in veracity judgments and sharing decisions among participants who made veracity judgments prior to sharing decisions, we submitted c scores of participants in the truth-prompt-present condition to a 2 (Judgment Type: veracity-judgment vs. sharing-decision) × 2 (Ideology-Congruence: congruent vs. incongruent) ANOVA with both factors varying within-subjects. Means and 95% confidence intervals in the four conditions are presented in Table 2. The ANOVA revealed a significant main effect of Ideology-Congruence, $F(1, 288) = 616.56$, $p < .001$, $\eta_p^2 = 0.682$, indicating that participants showed a lower acceptance threshold for ideology-congruent headlines compared to ideology-incongruent headlines. A significant main effect of Judgment Type further revealed that, although participants made veracity judgments immediately before they made a sharing decision, they still showed a higher acceptance threshold for sharing decisions compared to veracity judgments, $F(1, 288) = 288.14$, $p < .001$, $\eta_p^2 = 0.500$. These main effects were qualified by a significant two-way interaction between Judgment Type and Ideology-Congruence, $F(1, 288) = 57.46$, $p < .001$, $\eta_p^2 = 0.166$, indicating that partisan bias was more pronounced for veracity judgments, $F(1, 288) = 313.10$, $p < .001$, $\eta_p^2 = 0.521$, than sharing decisions, $F(1, 288) = 135.88$, $p < .001$, $\eta_p^2 = 0.321$.

Expanding on these analyses, we also submitted c scores for veracity judgments in the truth-prompt-present condition and c scores for sharing decisions in the truth-prompt-absent condition to a 2 (Judgment Type: veracity-judgment vs. sharing-decision) × 2 (Ideology-Congruence: congruent vs. incongruent) mixed ANOVA with the first factor varying between-subjects and the second varying within-subjects (corresponding to the main analysis in Experiment 1). Means and 95% confidence intervals in the four conditions are presented in Table 3. The ANOVA revealed a significant main effect of Ideology-Congruence, $F(1, 567) = 526.51$, $p < .001$, $\eta_p^2 = 0.481$, indicating that participants showed a lower acceptance threshold for ideology-congruent headlines compared to ideology-incongruent headlines. A significant main effect of

Table 3
Means and 95% Confidence Intervals of d′ *and* c *Scores for Veracity Judgments and Sharing Decisions as a Function of Ideology-Congruence (Ideology-Congruent vs. Ideology-Incongruent)*

	Veracity judgments prior to sharing decisions		Sharing decisions without prior veracity judgments	
SDT score	*M*	95% CI	*M*	95% CI
d′				
Ideology-congruent	0.580	[0.518, 0.643]	0.054	[−0.009, 0.118]
Ideology-incongruent	0.545	[0.482, 0.608]	0.007	[−0.058, 0.071]
C				
Ideology-congruent	−0.234	[−0.338, −0.131]	0.295	[0.191, 0.400]
Ideology-incongruent	0.567	[0.481, 0.652]	1.176	[1.089, 1.263]

Note. Between-participants comparison of scores by participants who reported veracity judgments prior to sharing decisions and participants who reported sharing decisions without prior veracity judgments, Experiment 4; SDT = signal detection theory.

Judgment Type further revealed that participants had a higher acceptance threshold for sharing decisions compared to veracity judgments, $F(1, 567) = 95.66$, $p < .001$, $\eta_p^2 = 0.144$. The two-way interaction of Ideology-Congruence and Judgment Type was not statistically significant, $F(1, 567) = 1.18$, $p = .279$, $\eta_p^2 = 0.002$.

Discussion

In addition to replicating the main findings of Experiment 1 (see also Experiments 2 and 3), Experiment 4 revealed two sets of noteworthy findings. First, consistent with Hypothesis 1, confirmatory analyses revealed that truth prompts increased truth sensitivity in sharing decisions. These findings provide further support for the idea that truth prompts may be an effective tool to increase the quality of information shared on social media (see Pennycook & Rand, 2022). However, as a caveat, it is worth noting that, although the current study used a rather heavy-handed truth-prompt manipulation (for alternatives, see Pennycook & Rand, 2022), the size of the obtained effect barely reached the benchmark for a small effect (see J. Cohen, 1988). The small size of this effect may also explain why some studies failed to obtain a significant effect of truth prompts with manipulations that are more subtle than the one used in the current study (e.g., Roozenbeek et al., 2021; see also Rathje et al., 2022). Although these issues do not necessarily undermine conclusions about the psychological underpinnings of truth sensitivity in sharing decisions, they should be considered for potential interventions to improve the quality of information shared on social media (see Pennycook & Rand, 2022).

Second, in addition to increasing truth sensitivity in sharing decisions, truth prompts also reduced partisan bias in sharing decisions. This finding is inconsistent with the null effect predicted under Hypothesis 2, which assumed that truth prompts would be effective in increasing truth sensitivity, but ineffective in reducing partisan bias. Interestingly, the effect size of the unexpected reduction in partisan bias was substantially larger compared to the relatively small increase in truth sensitivity. Whereas the predicted effect on truth sensitivity barely reached the benchmark of a small effect, the unexpected effect on partisan bias qualifies as medium in terms of current conventions (see J. Cohen, 1988). These results suggest that, although the impact of truth-prompt interventions on the overall quality of shared information may be relatively small overall, truth-prompt interventions may be quite effective in reducing partisan bias in information sharing. Beyond the unexpected effect on partisan bias, truth prompts increased participants' overall acceptance threshold in sharing decisions. This increase was greater for ideology-congruent than ideology-incongruent information, which led to the observed reduction in partisan bias. Put differently, truth prompts increased participants' threshold for sharing ideology-congruent information, which, in turn, reduced partisan bias in their sharing decisions.

Prediction of Misinformation Susceptibility

Expanding on debates about whether susceptibility to misinformation is better explained by partisan bias (Gawronski, 2021) or lack of truth sensitivity (Pennycook & Rand, 2021), we also conducted exploratory analyses investigating the extents to which belief in and sharing of false information are accounted for by truth sensitivity and partisan bias, respectively. Toward this end, we created indices of overall truth sensitivity and partisan bias (see Experiment 1), and then used the two indices as simultaneous predictors in multiple-regression analyses with acceptance of false information (i.e., false-alarm rates) as the criterion. The index of partisan-bias was created by calculating the difference between c scores for ideology-incongruent headlines and c scores for ideology-incongruent headlines, such that higher scores reflect a lower threshold for accepting ideology-congruent headlines compared to ideology-incongruent headlines (i.e., greater partisan bias). Mathematically, the calculation of partisan bias can be depicted with the following equation:

$$\begin{aligned} PB = {} & [-0.5 \times [z(H_{incongruent}) + z(FA_{incongruent})]] - [-0.5 \\ & \times [z(H_{congruent}) + z(FA_{congruent})]] \end{aligned} \tag{3}$$

which can be converted to

$$\begin{aligned} PB = {} & 0.5 \times [z(H_{congruent}) + z(FA_{congruent}) \\ & - z(H_{incongruent}) - z(FA_{incongruent})] \end{aligned} \tag{4}$$

An index of overall truth sensitivity was created by averaging d' scores for ideology-congruent and ideology-incongruent headlines, such that higher scores reflect greater sensitivity in distinguishing between true and false headlines. Mathematically, the calculation of overall truth sensitivity can be depicted with the following equation:

$$\begin{aligned} TS = {} & [[z(H_{congruent}) - z(FA_{congruent})] + [z(H_{incongruent}) \\ & - z(FA_{incongruent})]]/2 \end{aligned} \tag{5}$$

which can be converted to

$$\begin{aligned} TS = {} & 0.5 \times [z(H_{congruent}) - z(FA_{congruent}) + z(H_{incongruent}) \\ & - z(FA_{incongruent})] \end{aligned} \tag{6}$$

The converted equations for the calculation of overall truth sensitivity and partisan bias illustrate that the two indices are based on the same input data and equal treatment of data, the only difference being whether $z(FA_{congruent})$ and $z(H_{incongruent})$ enter the equation in an additive or subtractive manner. Whereas partisan bias increases with false-alarm rates for ideology-congruent headlines and decreases with hit rates for ideology-incongruent headlines, truth sensitivity decreases with false-alarm rates for ideology-congruent headlines and increases with hit rates for ideology-incongruent headlines. Across the four studies, overall truth sensitivity and partisan bias were largely uncorrelated, with correlations ranging from $r = -.022$ to .170 for veracity judgments and from $r = -.036$ to .125 for sharing decisions.

Because the criterion in our multiple-regression analyses (i.e., false-alarm rates) is used to calculate overall truth sensitivity and partisan bias, we ensured statistical independence of predictors and outcomes by calculating two scores for acceptance of false information, overall truth sensitivity, and partisan bias, respectively: one based on responses to headlines with odd item-numbers in our data set and one based on responses to headlines with even item-numbers. We then conducted two equivalent multiple-regression analyses with the data from each individual study. In the first analysis, we regressed acceptance of false headlines with odd item-numbers onto scores of overall truth sensitivity and partisan bias in responses to headlines

with even item-numbers. In the second analysis, we regressed acceptance of false headlines with even item-numbers onto scores of overall truth sensitivity and partisan bias in responses to headlines with odd item-numbers (essentially providing a cross-validation with the same data set). As partisan bias should increase susceptibility to ideology-congruent misinformation and decrease susceptibility to ideology-incongruent misinformation (Gawronski, 2021),[10] we conducted separate multiple-regression analyses for acceptance of ideology-congruent misinformation and acceptance of ideology-incongruent misinformation. Moreover, because belief in misinformation and sharing of misinformation may not necessarily be driven by the same factors, we conducted separate analyses for veracity judgments and sharing decisions, respectively. Expanding on the analyses of the data from the four individual studies, we repeated the same steps for the combined data from all four studies (see Curran & Hussong, 2009).

For veracity judgments (see Table 4), partisan bias reliably predicted belief in ideology-congruent and ideology-incongruent misinformation. Consistent with the argument that partisan bias should increase susceptibility to ideology-congruent misinformation and decrease susceptibility to ideology-incongruent misinformation (Gawronski, 2021), partisan bias showed a reliable positive relation to belief in ideology-congruent misinformation and a reliable negative relation to belief in ideology-incongruent misinformation. This pattern replicated in each individual study and with the combined data from all four studies regardless of whether responses to headlines with odd item-numbers were used to predict belief in false headlines with even item-numbers, or vice versa. Overall truth sensitivity showed a significant negative association with belief in ideology-incongruent misinformation in two of the four studies and the combined data from all four studies, but these associations were much smaller compared to the ones obtained for partisan bias. Overall truth sensitivity did not reliably predict belief in ideology-congruent misinformation. The only case where overall truth sensitivity showed a significant association with belief in ideology-congruent misinformation was the prediction of responses to headlines with odd item-numbers via responses to headlines with even item-numbers in the combined data set, but this relation did not replicate in the reverse prediction and in any of the four individual studies.

For sharing decisions (see Table 5), partisan bias reliably predicted sharing of ideology-congruent misinformation. This relation again replicated in each individual study and with the combined data from all four studies regardless of whether responses to headlines with odd item-numbers were used to predict sharing of false headlines with even item-numbers or vice versa. The pattern was less consistent for ideology-incongruent information, with partisan bias showing the expected negative association in five of the eight individual cases and in the combined data from the four studies. Overall truth sensitivity showed the expected negative relation to sharing of ideology-congruent misinformation in two of the eight individual cases and in one of the two cases in the combined data set, but there was also one individual case where truth sensitivity showed a significant positive association with sharing of ideology-congruent information (counter to the idea that greater truth sensitivity should reduce misinformation susceptibility). For sharing of ideology-incongruent information, overall truth sensitivity showed the expected negative relation in two of the eight individual cases, but these relations failed to reach statistical significance in the combined data set. Together, these results indicate that, although truth sensitivity and partisan bias were both associated with misinformation susceptibility, partisan bias was a stronger and much more reliable predictor of misinformation susceptibility than truth sensitivity.

General Discussion

The main goal of current research was to investigate truth sensitivity and partisan bias in responses to political (mis)information. Drawing on an SDT framework (see Batailler et al., 2022), we conceptualized *truth sensitivity* as the accurate discrimination between true and false information, and *partisan bias* as lower acceptance threshold for ideology-congruent information compared to ideology-incongruent information. Across four preregistered experiments, we examined (a) truth sensitivity and partisan bias in veracity judgments and decisions to share information and (b) determinants and correlates of truth sensitivity and partisan bias in responses to misinformation.

Veracity Judgments Versus Sharing Decisions

Across all four studies, truth sensitivity was greater for veracity judgments than sharing decisions. In fact, actual information veracity did not have any impact on participants' decision to share information online, in that truth sensitivity for sharing decisions did not significantly differ from chance level in all cases except one (i.e., sharing of ideology-congruent information after exposure to the truth prompt in Experiment 4). These results indicate that, although participants were clearly able to distinguish between true and false information, actual veracity did not matter for their decisions to share information online. This conclusion is consistent with prior findings suggesting that people often pay insufficient attention to veracity when they share information online (e.g., Pennycook, Epstein, et al., 2021; Pennycook et al., 2020). Interestingly, the obtained difference in truth sensitivity emerged even though participants were much more reluctant to share than accept information as true, as reflected in a higher acceptance threshold for sharing decisions as compared to veracity judgments. In other words, although participants were clearly very cautious in their decisions to share information (as reflected in the higher acceptance threshold for sharing decisions compared to veracity judgments), their greater caution did not increase the quality of the shared information in terms of information veracity (as reflected in the lower truth sensitivity for sharing decisions compared to veracity judgments). Across all four studies, participants also showed a substantially lower acceptance threshold for ideology-congruent information than ideology-incongruent information, providing strong evidence for partisan bias in both veracity judgments and sharing decisions. Counter to our expectation that partisan bias might be greater for sharing decisions than veracity judgments, partisan bias was not affected by judgment type in three of the four experiments despite sufficient statistical power to detect a small difference between judgment-type conditions; and in the only study that did obtain a significant difference, we found a pattern that was opposite to the predicted difference. Together, these results suggest that (a) truth sensitivity is greater for veracity judgments than sharing decisions, (b) acceptance

[10] Because partisan bias involves a general dismissal of all ideology-incongruent information, it should lead to a correct rejection of ideology-incongruent misinformation.

Table 4

Results of Multiple Regression Analyses Using Overall Truth Sensitivity and Partisan Bias to Predict Belief in Ideology-Congruent and Ideology-Incongruent Misinformation (Veracity Judgments)

	Belief in ideology-congruent misinformation				Belief in ideology-incongruent misinformation			
	Truth sensitivity		Partisan bias		Truth sensitivity		Partisan bias	
Experiment	β	*p*	β	*p*	β	*p*	β	*p*
Experiment 1								
Even–odd	−0.120	.097	0.195	.007	−0.024	.723	−0.372	<.001
Odd–even	0.016	.817	0.395	<.001	−0.104	.142	−0.284	<.001
Experiment 2								
Even–odd	−0.037	.450	0.356	<.001	−0.109	.017	−0.499	<.001
Odd–even	−0.008	.865	0.415	<.001	−0.141	.003	−0.404	<.001
Experiment 3								
Even–odd	0.050	.262	0.517	<.001	0.020	.694	−0.277	<.001
Odd–even	0.002	.967	0.497	<.001	0.017	.723	−0.313	<.001
Experiment 4								
Even–odd	−0.032	.567	0.285	<.001	−0.136	.009	−0.458	<.001
Odd–even	−0.014	.799	0.341	<.001	−0.128	.020	−0.359	<.001
Combined								
Even–odd	−0.066	.030	0.379	<.001	−0.085	.005	−0.401	<.001
Odd–even	−0.019	.513	0.440	<.001	−0.102	.001	−0.337	<.001

threshold is higher for sharing decisions than veracity judgments, and (c) partisan bias is strongly pronounced for both veracity judgments and sharing decisions.

Determinants of Truth Sensitivity

Consistent with prior research on the identification of fake news (e.g., Bago et al., 2020; Pennycook & Rand, 2019), we found that truth sensitivity in veracity judgments increased as a function of cognitive reflection (Experiment 2). However, because information veracity and source reliability are confounded in most studies on fake-news beliefs, it remains unclear if prior findings regarding the effect of cognitive reflection on fake-news susceptibility are driven by (a) enhanced ability to distinguish between true and false statements or (b) enhanced reliance on information about the sources' trustworthiness. In line with the idea that effects of cognitive reflection may be at least partly driven by information veracity, we found that cognitive reflection increased truth sensitivity in veracity judgments in the absence of source-related information. Whether the effect of cognitive reflection on truth sensitivity generalizes to sharing decisions remains unclear because our findings were inconclusive in this regard. Future research is needed to address this question.

Despite this ambiguity, our findings provide further support for the idea that truth prompts increase truth sensitivity in sharing decisions (Experiment 4). Although several studies have found that the quality of shared information increases when participants are prompted to think about the veracity of the focal information (see Pennycook & Rand, 2022), the reliability of such truth-prompt effects has been questioned by failed replications (Roozenbeek et al., 2021) and reanalyses of existing data suggesting that effects of truth prompts are

Table 5

Results of Multiple Regression Analyses Using Overall Truth Sensitivity and Partisan Bias to Predict Sharing of Ideology-Congruent and Ideology-Incongruent Misinformation (Sharing Decisions)

	Sharing of ideology-congruent misinformation				Sharing of ideology-incongruent misinformation			
	Truth sensitivity		Partisan bias		Truth sensitivity		Partisan bias	
Experiment	β	*p*	β	*p*	β	*p*	β	*p*
Experiment 1								
Even–odd	−0.033	.503	0.745	<.001	−0.178	.014	−0.108	.134
Odd–even	0.130	.014	0.708	<.001	0.108	.138	−0.144	.049
Experiment 2								
Even–odd	−0.016	.716	0.553	<.001	0.001	.979	−0.136	.009
Odd–even	0.060	.176	0.542	<.001	−0.016	.750	−0.148	.005
Experiment 3								
Even–odd	−0.116	.003	0.657	<.001	−0.072	.171	−0.044	.396
Odd–even	−0.075	.047	0.685	<.001	−0.130	.012	−0.080	.121
Experiment 4								
Even–odd	−0.025	.480	0.559	<.001	−0.019	.645	−0.152	<.001
Odd–even	−0.002	.964	0.550	<.001	−0.072	.084	−0.088	.035
Combined								
Even–odd	−0.053	.010	0.612	<.001	−0.045	.077	−0.114	<.001
Odd–even	0.015	.467	0.604	<.001	−0.041	.113	−0.106	<.001

limited to participants with liberal political ideology (Rathje et al., 2022). Although the current work used one of the more heavy-handed truth-prompt manipulations (see Pennycook & Rand, 2022) and the obtained effect size was rather small, our results support the effectiveness of truth prompts as a potential intervention to increase truth sensitivity in sharing decisions regardless of political affiliation.[11] Indeed, truth prompts even reduced partisan bias in sharing decisions, an unexpected finding discussed in the following section.

Determinants of Partisan Bias

Although cognitive reflection increased truth sensitivity in veracity judgments, cognitive reflection did not reduce partisan bias (Experiment 2). This result is consistent with the idea that that greater truth sensitivity does not necessarily reduce partisan bias. The independence of truth sensitivity and partisan bias is also supported by the results of correlational analyses, which revealed either no or a small positive association between truth sensitivity and partisan bias. Nevertheless, truth prompts effectively reduced partisan bias in sharing decisions (Experiment 4), counter to our prediction that truth prompts would increase truth sensitivity without affecting partisan bias. In fact, the effect of truth prompts on partisan bias was substantially larger than the obtained effect on truth sensitivity. Although the current work used one of the more heavy-handed truth-prompt manipulations (see Pennycook & Rand, 2022), these findings provide further support for the practical value of truth prompts as a potential intervention to increase the quality of information shared online, in that truth prompts may not only increase truth sensitivity but also reduce partisan bias in sharing decisions.

Despite this reassuring evidence, the psychological underpinnings of partisan bias are still unclear. On the one hand, it is possible that partisan bias arises from processes of motivated reasoning that aim to support and protect beliefs that are central to one's identity (Van Bavel & Pereira, 2018). On the other hand, partisan bias could be driven by non-motivational, cognitive processes that follow a pattern of Bayesian belief updating (Pennycook & Rand, 2021). Our preregistered confirmatory analyses regarding the impact of self-affirmation versus self-threat failed to provide compelling evidence for either of the two accounts (Experiment 3), in that self-affirmation (vs. self-threat) did not decrease partisan bias (as predicted by motivational accounts) or increase partisan bias (as predicted by cognitive accounts). Yet, the results of preregistered exploratory analyses provide preliminary support for cognitive accounts of partisan bias, in that partisan bias showed a significant positive association with self-confidence. A similar result was found in Experiment 1, where partisan bias showed a significant positive association with self-perceived ability in identifying news that is made up. Although the obtained associations were limited to veracity judgments in both studies, they suggest that partisan bias in veracity judgments increases with greater confidence, consistent with the assumptions of cognitive accounts. However, because the findings supporting this conclusion are merely correlational and the product of exploratory analyses, confirmatory tests using experimental approaches would help to further elucidate the psychological underpinnings of partisan bias in responses to misinformation.

Such tests would also be helpful to provide more conclusive evidence regarding the presumed contribution of motivational processes to partisan bias. Although the current findings provide no support for the idea that positive self-feelings may reduce partisan bias in sharing decisions, the results for partisan bias in veracity judgments remained inconclusive, prohibiting premature conclusions to either accept or reject the presumed contribution of motivational processes. Future research is needed to provide more compelling evidence in favor or against the idea that partisan bias in veracity judgments can be the product of motivational processes.

What Makes People Susceptible to Misinformation?

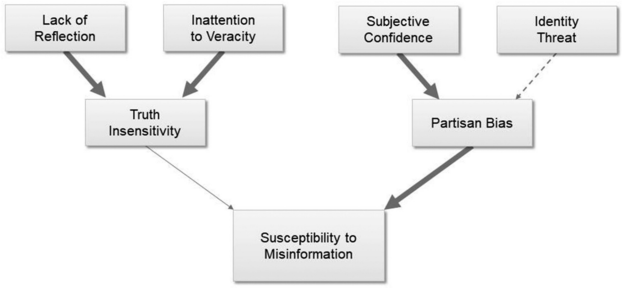
The current findings provide valuable insights for extant debates about why people fall for misinformation (Pennycook & Rand, 2021; Van Bavel & Pereira, 2018). From the perspective of SDT, the question of what makes people susceptible to misinformation can be restated as: *what makes people accept false information?* Or, stated with reference to the 2×2 matrix in Table 1: *what makes people prone to false alarms?* According to SDT, false alarms can be caused by low discrimination sensitivity or low response threshold (Green & Swets, 1966). Thus, if partisan bias in responses to political information is conceptualized as the difference in the acceptance threshold for ideology-congruent compared to ideology-incongruent information (see Batailler et al., 2022), a central question is whether misinformation susceptibility is explained by lack of truth sensitivity, partisan bias, or both. Our exploratory analyses of the data from the four studies revealed that partisan bias is a stronger and more reliable predictor of misinformation susceptibility than truth sensitivity (see Tables 4 and 5). For both veracity judgments and sharing decisions, greater levels of partisan bias were associated with stronger susceptibility to ideology-congruent misinformation and weaker susceptibility to ideology-incongruent misinformation. This pattern is consistent with the idea that partisan bias should (a) increase people's susceptibility to ideology-congruent misinformation by leading them to accept all ideology-congruent information regardless of veracity, but (b) decrease susceptibility to ideology-incongruent misinformation by leading them to reject all ideology-incongruent information regardless of veracity (Gawronski, 2021). The results further suggest that greater levels of truth sensitivity in veracity judgments are associated with weaker susceptibility to misinformation of either kind, but the obtained associations were unreliable across studies and the obtained effect sizes were substantially smaller compared to the ones obtained for partisan bias (see Tables 4 and 5).

The question of why people fall for misinformation can be conceptualized as a first level of analysis, in which misinformation susceptibility represents the phenomenon that needs to be explained, and truth sensitivity and partisan bias represent two explanatory constructs that may explain the focal phenomenon. Expanding on this first level of analysis, a potential follow-up question at a second level of analysis is what explains differences in truth sensitivity and partisan bias, respectively (see Figure 12). Going beyond patterns of behavioral responses, explanatory constructs at this second level of analysis may invoke mental processes that may underlie truth sensitivity and partisan bias (see Gawronski & Bodenhausen,

[11] The effect of truth prompts on truth sensitivity in Experiment 4 was not significantly qualified by participants' political affiliation (see the online supplemental materials). If anything, the effect of truth prompts was slightly stronger among Republicans compared to Democrats, different from the pattern obtained in Rathje et al.'s (2022) reanalysis of existing data.

Figure 12

Levels of Analysis in Understanding Susceptibility to Misinformation



Note. The first level of analysis treats misinformation susceptibility as a phenomenon that needs to be explained, and truth insensitivity and partisan bias as potential explanations for misinformation susceptibility. The second level of analysis treats truth insensitivity and partisan bias as phenomena that need to be explained, aiming to identify the mental underpinnings that explain truth insensitivity and partisan bias. The current findings suggest that, although truth insensitivity and partisan bias are both associated with misinformation susceptibility, partisan bias is a stronger and more reliable predictor of misinformation susceptibility than truth insensitivity. While truth insensitivity is caused by a lack of cognitive reflection and inattention to veracity, partisan bias increases with greater subjective confidence. The current findings are inconclusive regarding a potential link between identity threat and partisan bias.

2015). Our findings suggest that differences in truth sensitivity are at least partly due to differences in cognitive reflection (Experiment 2) and attention to veracity as an information property (Experiment 4). However, compared to partisan bias, truth sensitivity seems to play a less significant role for misinformation susceptibility. Partisan bias, on the other hand, seems to play a central role for misinformation susceptibility in both veracity judgments and sharing decisions, but the mental processes underlying partisan bias are still unclear. The results of our exploratory analyses suggest that partisan bias increases with subjective confidence, but further experimental work is needed to elucidate the psychological underpinnings of partisan bias in responses to misinformation. Such research is important not only for basic questions regarding the mental underpinnings of partisan bias; it is also important for interventions in applied contexts because the same intervention (e.g., self-affirmation) could either increase or decrease partisan bias depending on the mental processes underlying partisan bias.

A potential concern about the first level of analysis in our framework is that truth sensitivity and partisan bias are both based on false-alarm rates (i.e., misinformation susceptibility), raising questions about the independence of the phenomenon that needs to be explained (*explanandum*) and the constructs proposed to explain the focal phenomenon (*explanans*). There are two aspects of this argument that deserve attention: one involving statistical dependence and the other involving logical dependence. Statistical dependence is indeed an important issue, in that relations between two variables may be driven entirely by statistical overlap if the scores of one variable are calculated based on the scores of the other. In the current work, we circumvented this issue by using odd–even splits to create two non-overlapping subsets of data to calculate predictor and outcome scores. Regarding logical dependence, it is worth noting that any explanans (either by itself or in conjunction with auxiliary assumptions) must logically imply the explanandum; otherwise, it would not provide an explanation for the to-be-explained phenomenon. However, a major problem would arise if the to-be-explained phenomenon logically implies the construct proposed to explain the phenomenon (Hempel, 1970; for an example, see Greve, 2001). The latter is not the case in our framework because false alarms as the to-be-explained phenomenon do not logically imply specific levels of either truth sensitivity or partisan bias. Indeed, the fact that false-alarm rates can vary as a function of either truth sensitivity or partisan bias (or both) indicates that false-alarm rates alone do not provide any information about truth sensitivity and partisan bias. Thus, although the two explanatory constructs (i.e., truth sensitivity, partisan bias) logically imply the to-be-explained phenomenon (i.e., false alarms) by virtue of their explanatory relation, the reverse is not the case, which addresses potential concerns about logical dependence.

Conflicting Claims About Partisan Bias

The current findings stand in stark contrast to conclusions by Pennycook and Rand (2021) that partisan bias does not explain susceptibility to misinformation (see also Pennycook & Rand, 2019). How can the current findings be reconciled with these claims? The main reason for the apparent conflict is that Pennycook and Rand based their conclusions on a conceptualization of partisan bias in terms of truth discernment, which is fundamentally different from the current conceptualization in terms of acceptance thresholds (see Gawronski, 2021). According to Pennycook and Rand, partisan bias should lead to lower truth discernment (i.e., lower truth sensitivity in terms of SDT) for ideology-congruent compared to ideology-incongruent information. Yet, as pointed out by Pennycook and Rand, the available evidence suggests the opposite, in that truth discernment is higher, not lower, for ideology-congruent compared to ideology-incongruent information. This finding led them to dismiss partisan bias as a factor contributing to misinformation susceptibility.

Although debates about definitions of psychological constructs are notoriously difficult to resolve, there are strong arguments in favor of the current conceptualization, stipulating that partisan bias involves a lower acceptance threshold for ideology-congruent information compared to ideology-congruent information (see Batailler et al., 2022; Gawronski, 2021). Indeed, when partisan bias is conceptualized in this manner, there is substantial evidence for partisan bias in prior studies, including Pennycook and Rand's own work (see the results of reanalyses by Batailler et al., 2022; Gawronski, 2021). Different from the current label *partisan bias*, Pennycook and Rand (2021) describe the phenomenon as an effect of ideology-congruence on overall belief. Yet, regardless of terminological preferences, the data that led Pennycook and Rand to dismiss "partisan bias" as an explanation for misinformation susceptibility are perfectly consistent with the current findings, all of which suggest a greater willingness to accept ideology-congruent information compared to ideology-incongruent information. Importantly, the current findings suggest that this ubiquitous tendency (regardless of whether it is called *partisan bias* or something else) plays a central role for misinformation susceptibility—and, in fact, a much greater role than truth sensitivity.

Theoretical Implications

Our conclusions about the significance of partisan bias in responses to misinformation are consistent with a broad range of prior research on motivated skepticism (Ditto & Lopez, 1992), wishful thinking (Kruglanski et al., 2020), political tribalism (Finkel et al., 2020), cognitive responses in persuasion (Petty & Cacioppo, 1986), defensive processing of counterattitudinal information (Chaiken et al., 1989), social identity and information processing (Van Bavel & Pereira, 2018), partisan bias in recollective memory (Calvillo et al., 2022), and biased belief updating (Tappin et al., 2017). Yet, the current findings are inconclusive about the extent to which partisan bias in veracity judgments arises from a desire to support and protect identity-related beliefs, as suggested by motivational accounts (Van Bavel & Pereira, 2018). We obtained much stronger evidence for cognitive accounts, suggesting that partisan bias in veracity judgments is a product of high subjective confidence about the accuracy of one's beliefs (Pennycook & Rand, 2021). The latter idea suggests interesting conceptual links to research on illusory-truth effects (see Brashier & Marsh, 2020; Schwarz et al., 2007), in that repetition of ideology-congruent information in echo chambers may increase the fluency of processing ideology-congruent information, which, in turn, may contribute to partisan bias in veracity judgments by increasing people's confidence about the accuracy of their ideological beliefs.

Beyond their value for the question of why people fall for misinformation, the current findings also have significant theoretical implications for research on truth judgments (for a review, see Brashier & Marsh, 2020). The large partisan bias obtained in the current studies is consistent with the notion that cognitive consistency plays a central role in naïve assessments of veracity (e.g., Gawronski, 2012; Higgins, 2012; Schwarz & Jalbert, 2020). Research suggests that people are more likely to judge information as true when it is consistent with their beliefs than when it is inconsistent with their beliefs (Brashier & Marsh, 2020). According to extant theories of cognitive consistency, inconsistency functions as a cue for potential errors in one's system of beliefs that would need to be corrected (Gawronski, 2012; Gawronski & Brannon, 2019). Unless it is possible to resolve inconsistency between new information and one's prior beliefs via additional information that permits both to be true, the inconsistency must be resolved either by rejecting the new information as false or by updating one's beliefs in line with the new information (Pinquart et al., 2021). Research on the role of mental models in the assessment of (in) consistency suggests that, in such cases, people are more likely to "explain away" the new information than to update their prior beliefs (Johnson-Laird, 2012). Dissonance theory further suggests that this tendency should be more pronounced for strongly held beliefs compared to weakly held beliefs (Festinger, 1957), consistent with the current finding that partisan bias was positively associated with subjective confidence. In the current studies, participants showed a strong tendency to resolve conflicts between their ideological beliefs and ideology-incongruent news headlines by dismissing the news headlines as false, and this tendency increased as a function of subjective confidence. Thus, although partisan bias in veracity judgments seems highly problematic for its potential to make people susceptible to misinformation, it can be understood as the product of basic processes in naïve assessments of veracity (see Brashier & Marsh, 2020).

Although these considerations suggest a potential mechanism underlying partisan bias in veracity judgments, the mechanisms underlying truth sensitivity in veracity judgments are still unclear. The current research suggests that more thoughtful processing increases truth sensitivity in veracity judgments (Experiment 2). However, it remains unclear what specific cues participants relied on that made them better in distinguishing between true and false headlines. As prior research on this question always presented false statements with a dubious news source and true statements with a mainstream news source (e.g., Bago et al., 2020), earlier findings are consistent with the idea that cognitive reflection might increase the use of cues about the source's trustworthiness (see Schwarz & Jalbert, 2020). However, such an explanation does not apply to the current finding that cognitive reflection increased truth sensitivity in the absence of source information. Greater reliance on inconsistency with prior beliefs also does not explain the observed effect of cognitive reflection on truth sensitivity, because greater reliance on inconsistency as a cue should increase partisan bias, not truth sensitivity. Thus, although enhanced use of source-related information and inconsistency with prior beliefs can be ruled out as explanations for higher levels of truth sensitivity in the current studies, the mechanisms underlying truth sensitivity in veracity judgments are still unclear. Future research is needed to address this question.

Regarding the mechanisms underlying sharing decisions, it seems reasonable to assume that naïve assessments of veracity have downstream effects on people's decisions to share information. This conclusion is consistent with the findings of Experiment 4, showing that prior judgments of veracity influenced both truth sensitivity and partisan bias in sharing decisions. However, the current findings also suggest that perceptions of veracity are not the only factor that influence sharing decisions. Clearly, people do not share everything they believe to be true, and sometimes people share information even when they know that it is false (Effron & Raj, 2020). In the truth-prompt-present condition of Experiment 4, the disconnect between perceptions of veracity and sharing decisions was reflected in the finding that prior veracity judgments did not fully match subsequent sharing decisions, in that (a) participants showed a substantially higher threshold in their sharing decisions compared to their prior veracity judgments of the same headlines and (b) truth sensitivity was substantially lower for sharing decisions compared to prior veracity judgments of same headlines. These differences cannot be explained by insufficient attention to accuracy (e.g., Pennycook, Epstein, et al., 2021; Pennycook et al., 2020) because participants in the truth-prompt-present condition judged the veracity of every headline immediately before making a sharing decision for that headline. A more plausible explanation is that veracity judgments and sharing decisions are shaped by different goals. For example, whereas veracity judgments are likely shaped by accuracy goals, sharing decisions are presumably influenced by various additional goals, including goals related to the expression of one's social identity (Chaiken et al., 1989; Van Bavel & Pereira, 2018). Thus, although the current findings suggest that truth nudges can be helpful to increase truth sensitivity and reduce partisan bias in sharing decisions, our findings also indicate that an exclusive focus on accuracy goals is insufficient to provide a full understanding of why people share (true and false) information.

Open Questions and Limitations

Although the current research provides valuable insights into why people fall for misinformation, several questions remain unanswered, calling for more research in this area. First, although we attempted to provide deeper insights into the mental processes

underlying partisan bias, the psychological underpinnings of partisan bias are still unclear, partly because our preregistered confirmatory analyses failed to obtain reliable effects of self-affirmation versus self-threat (see also Lyons et al., 2022). Results of our preregistered exploratory analyses suggest that partisan bias in responses to misinformation is associated with greater subjective confidence. However, based on this association alone, it remains unclear whether (a) greater confidence causes partisan bias, (b) a tendency to show partisan bias increases confidence, or (c) the association between confidence and partisan bias is driven by a third factor. Moreover, because the obtained association between partisan bias and subjective confidence is the product of exploratory correlational analyses, confirmatory experimental research would be helpful to provide more compelling evidence for the psychological underpinnings of partisan bias.

Another important question for future research concerns the respective roles of source and content in responses to misinformation. A substantial amount of work in this area is concerned with susceptibility to fake news (for a review, see Pennycook & Rand, 2021), defined as "fabricated information that mimics news media content in form but […] lack(s) the news media's editorial norms and processes for ensuring the accuracy and credibility of information" (Lazer et al., 2018, p. 1094). An important methodological aspect of research on fake-news beliefs is that it involves a paradigmatic confound of information veracity and source reliability, in that false statements are always presented with a dubious news source and true statements are always presented with mainstream news source (for two notable exceptions, see Pehlivanoglu et al., 2021; Traberg & van der Linden, 2022). The current research aimed to overcome this limitation by focusing exclusively on effects of content-related features, namely information veracity and political slant. However, the elimination of source information in the current studies raises the question of how source-related and content-related features may jointly influence responses to misinformation. Two important source characteristics in this regard are (a) the perceived political leaning of a given source independent of its perceived reliability and (b) the perceived reliability of a given source independent of its perceived political leaning. Future research manipulating different source characteristics in addition to content-related features would help to further elucidate the unique and interactive roles of source and content in responses to misinformation.

It also seems appropriate to acknowledge two methodological limitations of the employed measure of sharing decisions. First, although the measure is widely used in research on misinformation susceptibility, it could be criticized for capturing hypothetical decisions without real-world consequences. We generally agree with this concern. Yet, it is worth noting that self-reported willingness to share political news in online surveys has been found to be strongly associated with actual news sharing on social media (Mosleh et al., 2020). Thus, despite its obvious limitations, the measure has at least some validity for understanding actual sharing decisions. Second, the measure could be criticized for not capturing participants' motivations behind their sharing decisions. For example, people may sometimes share information online to correct it or to express their disapproval. Although the current studies do not provide any data on the possibility of corrective motivation, the large partisan-bias effects in the current studies speak against the idea of sharing for the expression of disapproval. Yet, because the criticisms apply to virtually all research on the sharing of misinformation, future studies on the motivations behind sharing decisions would be helpful to gain deeper insights into the mental underpinnings of sharing decisions.

Constraints on Generality

Another important issue concerns the generalizability of the current findings to other populations and stimulus materials. Although the samples in the current studies were relatively diverse in terms of their demographic characteristics and political leanings, all four experiments were conducted with participants from the United States with two sets of political headlines. This limitation raises two important questions about the generalizability of the current findings. First, the current political climate in the United States is characterized by an extreme level of polarization that seems less common in many other parts of the world (Finkel et al., 2020). In addition, the United States is unique for its two-party system, which makes it different from other democratic nations with multiparty systems. These differences raise the question of whether the current findings generalize to countries other than the United States. This question seems especially important for our findings regarding the role of partisan bias, which might be attenuated in less polarized democracies with multi-party systems or in nondemocratic societies. Second, although the set of stimuli in the current studies was considerably larger than the stimulus sets in prior studies in this area (e.g., Pennycook & Rand, 2019), it would be desirable to replicate the current findings with other stimulus sets to demonstrate their stimulus independence. Going beyond the current focus on political (mis)information, it would also be valuable to obtain evidence for conceptually similar effects in other domains, such as (mis)information about COVID-19 or vaccines. Although there is no direct translation of partisan bias to these domains, it is still possible to analyze differences in acceptance thresholds for different kinds of information (e.g., pro-vaccine vs. anti-vaccine) and investigate their determinants, consequences, and individual-difference correlates. Based on the current finding that partisan bias played a more significant role in misinformation susceptibility than truth sensitivity, research of this kind could help to provide a better understanding of why people fall for misinformation more broadly.

Concerns could also be raised about the thorough screening of research materials in the current work (see Appendix), which may similarly undermine the generalizability of the obtained results. Careful prescreening is essential for research on truth sensitivity and partisan bias because both constructs require valid operationalizations of actual veracity and political slant (e.g., some headlines may be perceived as partisan by Democrats but not by Republicans, or vice versa). Without careful prescreening, the results of SDT analyses can be distorted, leading to inaccurate conclusions. Nevertheless, an interesting question that cannot be addressed with such restricted materials is how people respond to different kinds of political headlines more broadly, and which particular features influence their veracity judgments and sharing decisions. Future research may help to address this question.

A final issue concerns the suitability of the current stimulus materials for future studies that aim to replicate the current findings. Although we deem replications with identical materials valuable to determine the reliability of an observed effect (see Gawronski, 2022; Gawronski & Brannon, 2021; Gawronski et al., 2017, 2018,

2022), an important issue in research on misinformation susceptibility is the time-sensitivity of the stimulus materials (see Pennycook, Binnendyk, et al., 2021). While some of the headlines in our studies may still evoke similar responses at the time when this article is published, other headlines may be outdated by that time—and most, if not all, headlines will likely be outdated a few years later. Whether replication attempts with outdated stimulus materials will be successful is an open question, but we would deem it more appropriate to conduct replication studies with new stimulus materials that are (a) timely while the data are collected and (b) identified based on the same rigorous criteria used in the current studies (see Appendix). This approach should also be followed for conceptual replication studies in other content domains (e.g., Covid-19 vaccines) and with other populations (e.g., samples from other countries).

Conclusion

Drawing on a signal-detection framework, the current research investigated two distinct aspects of misinformation susceptibility: *truth sensitivity*, conceptualized as the accurate discrimination between true and false information, and *partisan bias*, conceptualized as lower acceptance threshold for ideology-congruent information compared to ideology-incongruent information. Across four preregistered experiments, we found that decisions to share information online were largely unaffected by actual information veracity, although participants were able to distinguish between true and false information to a considerable extent. Moreover, a strong partisan bias emerged for both veracity judgments and sharing decisions, with partisan bias being unrelated to the overall degree of truth sensitivity. While truth sensitivity increased as a function of cognitive reflection during encoding, partisan bias increased as a function of subjective confidence. Truth sensitivity and partisan bias were both associated with misinformation susceptibility, but partisan bias was a stronger and more reliable predictor of misinformation susceptibility than truth sensitivity. Together, these findings provide valuable insights for extant debates about why people fall for misinformation. Given the significance of partisan bias for understanding susceptibility to misinformation, more research is needed to clarify the psychological underpinnings of partisan bias.

References

Alicke, M. D., & Govorun, O. (2005). The better-than-average effect. In M. D. Alicke, D. A. Dunning, & J. I. Krueger (Eds.), *The self in social judgment* (pp. 85–106). Psychology Press.

Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology*, *20*(1), 1–48. https://doi.org/10.1080/10463280802613866

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General*, *149*(8), 1608–1613. https://doi.org/10.1037/xge0000729

Batailler, C., Brannon, S. M., Teas, P. E., & Gawronski, B. (2022). A signal detection approach to understanding the identification of fake news. *Perspectives on Psychological Science*, *17*(1), 78–98. https://doi.org/10.1177/1745691620986135

Brandt, M. J., & Sleegers, W. W. (2021). Evaluating belief system networks as a theory of political belief system dynamics. *Personality and Social Psychology Review*, *25*(2), 159–185. https://doi.org/10.1177/1088868321993751

Brashier, N. M., & Marsh, E. J. (2020). Judging truth. *Annual Review of Psychology*, *71*(1), 499–515. https://doi.org/10.1146/annurev-psych-010419-050807

Briñol, P., & Petty, R. E. (2022). Self-validation theory: An integrative framework for understanding when thoughts become consequential. *Psychological Review*, *129*(2), 340–367. https://doi.org/10.1037/rev0000340

Calvillo, D. P., Harris, J. D., & Hawkins, W. C. (2022). Partisan bias in false memories for misinformation about the 2021 US Capitol riot. *Memory*. Advance online publication. https://doi.org/10.1080/09658211.2022.2127771

Chaiken, S., Liberman, A., & Eagly, A. H. (1989). Heuristic and systematic processing within and beyond the persuasion context. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 212–252). Guilford Press.

Cohen, G. L., Sherman, D. K., Bastardi, A., Hsu, L., McGoey, M., & Ross, L. (2007). Bridging the partisan divide: Self-affirmation reduces ideological closed-mindedness and inflexibility in negotiation. *Journal of Personality and Social Psychology*, *93*(3), 415–430. https://doi.org/10.1037/0022-3514.93.3.415

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, *14*(2), 81–100. https://doi.org/10.1037/a0015914

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, *63*(4), 568–584. https://doi.org/10.1037/0022-3514.63.4.568

Dunning, D. (2011). The Dunning–Kruger effect: On being ignorant of one's own ignorance. *Advances in Experimental Social Psychology*, *44*, 247–296. https://doi.org/10.1016/B978-0-12-385522-0.00005-6

Effron, D. A., & Raj, M. (2020). Misinformation and morality: Encountering fake-news headlines makes them seem less unethical to publish and share. *Psychological Science*, *31*(1), 75–87. https://doi.org/10.1177/0956797619887896

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Festinger, L. (1957). *A theory of cognitive dissonance*. Row Peterson.

Finkel, E. J., Bail, C. A., Cikara, M., Ditto, P. H., Iyengar, S., Klar, S., Mason, L., McGrath, M. C., Nyhan, B., Rand, D. G., Skitka, L. J., Tucker, J. A., Van Bavel, J. J., Wang, C. S., & Druckman, J. N. (2020). Political sectarianism in America. *Science*, *370*(6516), 533–536. https://doi.org/10.1126/science.abe1715

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Gawronski, B. (2012). Back to the future of dissonance theory: Cognitive consistency as a core motive. *Social Cognition*, *30*(6), 652–668. https://doi.org/10.1521/soco.2012.30.6.652

Gawronski, B. (2021). Partisan bias in the identification of fake news. *Trends in Cognitive Sciences*, *25*(9), 723–724. https://doi.org/10.1016/j.tics.2021.05.001

Gawronski, B. (2022). Attitudinal effects of stimulus co-occurrence and stimulus relations: Paradoxical effects of cognitive load. *Personality and Social Psychology Bulletin*, *48*(10), 1438–1450. https://doi.org/10.1177/01461672211044322

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hütter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of Personality and Social Psychology*, *113*(3), 343–376. https://doi.org/10.1037/pspa0000086

Gawronski, B., & Bodenhausen, G. V. (2015). Social-cognitive theories. In B. Gawronski & G. V. Bodenhausen (Eds.), *Theory and explanation in social psychology* (pp. 65–83). Guilford Press.

Gawronski, B., & Brannon, S. M. (2019). What is cognitive consistency and why does it matter? In E. Harmon-Jones (Ed.), *Cognitive dissonance: Reexamining a pivotal theory in psychology* (2nd ed., pp. 91–116). American Psychological Association.

Gawronski, B., & Brannon, S. M. (2021). Attitudinal effects of stimulus co-occurrence and stimulus relations: Range and limits of intentional control. *Personality and Social Psychology Bulletin*, *47*(12), 1654–1667. https://doi.org/10.1177/0146167220982906

Gawronski, B., Brannon, S. M., & Ng, N. L. (2022). Debunking misinformation about a causal link between vaccines and autism: Two preregistered tests of dual-process versus single-process predictions (with conflicting results). *Social Cognition*, *40*(6), 580–599. https://doi.org/10.1521/soco.2022.40.6.580

Gawronski, B., Conway, P., Armstrong, J., Friesdorf, R., & Hütter, M. (2018). Effects of incidental emotions on moral dilemma judgments: An analysis using the CNI model. *Emotion*, *18*(7), 989–1008. https://doi.org/10.1037/emo0000399

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.

Greve, W. (2001). Traps and gaps in action explanation: Theoretical problems of a psychology of human action. *Psychological Review*, *108*(2), 435–451. https://doi.org/10.1037/0033-295X.108.2.435

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. Presidential election. *Science*, *363*(6425), 374–378. https://doi.org/10.1126/science.aau2706

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1), Article eaau4586. https://doi.org/10.1126/sciadv.aau4586

Hempel, C. G. (1970). *Aspects of scientific explanation and other essays in the philosophy of science*. Free Press.

Higgins, E. T. (2012). *Beyond pleasure and pain: How motivation works*. Oxford University Press.

Johnson-Laird, P. N. (2012). Mental models and consistency. In B. Gawronski & F. Strack (Eds.), *Cognitive consistency: A fundamental principle in social cognition* (pp. 225–244). Guilford Press.

Kruglanski, A. W., Jasko, K., & Friston, K. (2020). All thinking is 'wishful' thinking. *Trends in Cognitive Sciences*, *24*(6), 413–424. https://doi.org/10.1016/j.tics.2020.03.004

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news: Addressing fake news requires a multidisciplinary effort. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. https://doi.org/10.1177/1529100612451018

Liberman, A., & Chaiken, S. (1992). Defensive processing of personally relevant health messages. *Personality and Social Psychology Bulletin*, *18*(6), 669–679. https://doi.org/10.1177/0146167292186002

Lyons, B. A., Farhart, C. E., Hall, M. P., Kotcher, J., Levendusky, M., Miller, J. M., Nyhan, B., Raimi, K. T., Reifler, J., Saunders, K. L., Skytte, R., & Zhao, X. (2022). Self-affirmation and identity-driven political behavior. *Journal of Experimental Political Science*, *9*(2), 225–240. https://doi.org/10.1017/XPS.2020.46

Lyons, B. A., Montgomery, J. M., Guess, A. M., Nyhan, B., & Reifler, J. (2021). Overconfidence in news judgments is associated with false news susceptibility. *Proceedings of the National Academy of Sciences*, *118*(23), Article e2019527118. https://doi.org/10.1073/pnas.2019527118

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Taylor and Francis.

McQueen, A., & Klein, W. M. P. (2006). Experimental manipulations of self-affirmation: A systematic review. *Self and Identity*, *5*(4), 289–354. https://doi.org/10.1080/15298860600805325

Mosleh, M., Pennycook, G., Arechar, A. A., & Rand, D. G. (2021). Cognitive reflection correlates with behavior on Twitter. *Nature Communications*, *12*(1), Article 921. https://doi.org/10.1038/s41467-020-20043-0

Mosleh, M., Pennycook, G., & Rand, D. G. (2020). Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter. *PLoS ONE*, *15*(2), Article e0228882. https://doi.org/10.1371/journal.pone.0228882

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Nelson, J. L., & Taneja, H. (2018). The small, disloyal fake news audience: The role of audience availability in fake news consumption. *New Media & Society*, *20*(10), 3720–3737. https://doi.org/10.1177/1461444818758715

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Pehlivanoglu, D., Lin, T., Deceus, F., Heemskerk, A., Ebner, N. C., & Cahill, B. S. (2021). The role of analytical reasoning and source credibility on the evaluation of real and fake full-length news articles. *Cognitive Research: Principles and Implications*, *6*(1), Article 24. https://doi.org/10.1186/s41235-021-00292-3

Pennycook, G. (2023). A framework for understanding reasoning errors: From fake news to climate change and beyond. *Advances in Experimental Social Psychology*. Advance online publication. https://doi.org/https://doi.org/10.1016/bs.aesp.2022.11.003

Pennycook, G., Binnendyk, J., Newton, C., & Rand, D. G. (2021). A practical guide to doing behavioral research on fake news and misinformation. *Collabra: Psychology*, *7*(1), Article 25293. https://doi.org/10.1525/collabra.25293

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595. https://doi.org/10.1038/s41586-021-03344-2

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, *31*(7), 770–780. https://doi.org/10.1177/0956797620939054

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, *25*(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*, *13*(1), Article 2333. https://doi.org/10.1038/s41467-022-30073-5

Petty, R. E., & Cacioppo, J. T. (1986). The Elaboration Likelihood Model of persuasion. *Advances in Experimental Social Psychology*, *19*, 123–205. https://doi.org/10.1016/S0065-2601(08)60214-2

Pinquart, M., Endres, D., Teige-Mocigemba, S., Panitz, C., & Schütz, A. C. (2021). Why expectations do or do not change after expectation violation:

A comparison of seven models. *Consciousness and Cognition*, *89*, Article 103086. https://doi.org/10.1016/j.concog.2021.103086

Rathje, S., Roozenbeek, J., Traberg, C. S., Van Bavel, J., & Van der Linden, S. (2022). Letter to the Editors of Psychological Science: Meta-analysis reveals that accuracy nudges have little to no effect for U.S. conservatives: Regarding Pennycook et al. (2020). *Psychological Science*.

Roozenbeek, J., Freeman, A. L. J., & Van der Linden, S. (2021). How accurate are accuracy-nudge interventions? A preregistered direct replication of Pennycook et al. (2020). *Psychological Science*, *32*(7), 1169–1178. https://doi.org/10.1177/09567976211024535

Ross, R. M., Rand, D. G., & Pennycook, G. (2021). Beyond "fake news": Analytic thinking and the detection of false and hyperpartisan news headlines. *Judgment and Decision Making*, *16*(2), 484–504. https://doi.org/10.1017/S1930297500008640

Schwarz, N., & Jalbert, M. (2020). When (fake) news feels true: Intuitions of truth and the acceptance and correction of misinformation. In R. Greifeneder, M. Jaffé, E. J. Newman, & N. Schwarz (Eds.), *The psychology of fake news: Accepting, sharing, and correcting misinformation* (pp. 73–90). Routledge.

Schwarz, N., Sanna, L. J., Skurnik, I., & Yoon, C. (2007). Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and public information campaigns. *Advances in Experimental Social Psychology*, *39*, 127–161. https://doi.org/10.1016/S0065-2601(06)39003-X

Sherman, D. K., & Cohen, G. L. (2006). The psychology of self-defense: Self-affirmation theory. *Advances in Experimental Social Psychology*, *38*, 183–242. https://doi.org/10.1016/S0065-2601(06)38004-5

Sherman, D. K., Nelson, L. D., & Steele, C. M. (2000). Do messages about health risks threaten the self? Increasing the acceptance of threatening health messages via self-affirmation. *Personality and Social Psychology Bulletin*, *26*(9), 1046–1058. https://doi.org/10.1177/01461672002611003

Stanislav, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, and Computers*, *31*(1), 137–149. https://doi.org/10.3758/BF03207704

Tappin, B. M., Pennycook, G., & Rand, D. G. (2020). Bayesian or biased? Analytic thinking and political belief updating. *Cognition*, *204*, Article 104375. https://doi.org/10.1016/j.cognition.2020.104375

Tappin, B. M., Van Der Leer, L., & McKay, R. T. (2017). The heart trumps the head: Desirability bias in political belief revision. *Journal of Experimental Psychology: General*, *146*(8), 1143–1149. https://doi.org/10.1037/xge0000298

Traberg, C. S., & van der Linden, S. (2022). Birds of a feather are persuaded together: Perceived source credibility mediates the effect of political bias on misinformation susceptibility. *Personality and Individual Differences*, *185*, Article 111269. https://doi.org/10.1016/j.paid.2021.111269

Trafimow, D. (2015). Rational actor theories. In B. Gawronski & G. V. Bodenhausen (Eds.), *Theory and explanation in social psychology* (pp. 245–265). Guilford Press.

Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. *Trends in Cognitive Sciences*, *22*(3), 213–224. https://doi.org/10.1016/j.tics.2018.01.004

(*Appendices follow*)

Appendix

Procedure for the Selection of Headlines Used in the Current Studies

Headline Search

We began the selection process by searching through mainstream news websites and fact-checking websites that report on misinformation. For true information, news sources that we utilized were *CNN*, *Fox News*, *BBC News*, *AP News*, *New York Times*, and *Politico*. For false information, we used FactCheck.org, Snopes, LeadStories, and Politifact. In addition to following best practices for the selection of stimulus materials for research on misinformation (see Pennycook, Binnendyk, et al., 2021), our search was guided by two heuristic criteria. First, we searched for headlines that both conservatives and liberals would consensually identify as either pro-Democrat or pro-Republican. Thus, the focal issue had to be clearly partisan and directly related to one of the two parties instead of having a distal link to the two parties (e.g., headlines about COVID-19 or Black Lives Matter). Second, the headlines had to be suitable for several months in the future for the active study. Hence, headlines could not rely too heavily on the context of the time in which it was published or be easily contradicted in the future. We regularly checked the identified websites from September 2020 to March 2021 and added headlines that met our inclusion criteria to a shared data base. For each headline added to the data base, we included the following information: (a) headline in its original wording, (b) veracity of the headline, (c) political leaning of the headline, (d) original source of the headline, (e) publication date of the headline, (f) the date we identified the headline, (g) the context of the headline specifying its truth or falsity, (h) information on fact-checking sources, (i) the initials of the person who added the headline to the data base, and (j) other notes on the headline.

Headline Screening

In an initial screening of the identified headlines, we excluded headlines whose partisanship might be perceived differently among Democrats and Republicans (e.g., a headline may be perceived as pro-Democrat among Democrats, but as relatively neutral among Republicans). We also excluded headlines that attributed a statement to a person because such headlines would lead to ambiguity about what participants are supposed to judge in the main study (e.g., for a headline stating *Person A said XYZ*, the veracity question in the main study could be interpreted as asking if XYZ is true or if it is true that Person A said XYZ). Finally, we excluded true headlines whose content seemed too widely known.

To further narrow down the shortlist of headlines for our pilot study, we then re-examined our initial short-list to determine whether the headlines were still relevant and not outdated or otherwise time sensitive. In addition, we rescreened the headlines for sufficiently strong partisanship. We further excluded headlines whose veracity seemed ambiguous, along with headlines whose wording seemed confusing. Lastly, we excluded headlines that were contradictory with other headlines on our shortlist.

Pilot Study 1

In our first pilot study, we recruited 120 self-identified conservatives and 120 self-identified liberals via CloudResearch and asked them two questions for each of 120 preselected headlines: (a) *How would you rate the political slant of this statement?* (b) *Have you heard about the claim in this statement before?* Responses to the first question were measured with an unnumbered 7-point rating scale with the endpoints *Very Pro-Democrat* (recorded as 1) and *Very Pro-Republican* (recorded as 7). Responses to the second question were measured with an unnumbered 7-point rating scale with the endpoints *Very confident I did not hear this before* (recorded as 1) and *Very confident I did hear this before* (recorded as 7). Based on the collected pilot data, we calculated the means and modes of partisanship scores and average familiarity scores for each headline. Scores were calculated separately for self-identified conservatives and self-identified liberals. Based on the obtained scores, we first eliminated headlines that were insufficiently partisan. Toward this end, we excluded all pro-Democrat headlines with a mean partisanship score > 3.00 among either liberals or conservatives, and all pro-Republican headlines with a mean partisanship score < 5.00 among either liberals or conservatives. Next, we eliminated headlines with a mode partisanship score of 4 among either liberals or conservatives. After applying these criteria, the list of headlines identified in our first pilot study included 19 true pro-Democrat headlines, 15 true pro-Republican headlines, 19 false pro-Democrat headlines, and 32 false pro-Republican headlines.

We then narrowed down the lists for each of the four categories to 15 headlines, which was the number of headlines in the smallest category (i.e., 15 true pro-Republican headlines). Toward this end, we matched the lists of true pro-Democrat headlines and true pro-Republican headlines by familiarity. The same was done for the lists of false pro-Democrat headlines and false pro-Republican headlines. We tried to match the headlines as much as possible by their familiarity scores, as determined by the politically congruent and politically incongruent groups, as well as by the difference between these two scores. Thus, each true pro-Democrat headline would have a true pro-Republican headline that would be equally familiar to both Democrats and Republicans, and vice versa. Using this procedure, we first eliminated four headlines from the list of true pro-Democrat headlines and four headlines from the list of false pro-Democrat headlines, leaving us with 15 headlines in each of the two categories. We then matched the false pro-Republican headlines with the false pro-Democrat headlines in terms of familiarity and eliminated 17 pro-Republican headlines that were not congruent on familiarity scores. This left us with a final list of 15 headlines per category. The final list of headlines and the pilot data for the selection of headlines are available at https://osf.io/d2rne/. The OSF page also includes a data file for the final set of headlines that includes information on: (a) the veracity of the headline, (b) the political leaning of the headline, (c) the broader background of the headline, (d) context information that helped us determine the veracity of the headline, (e) the original source of the headline, (f) publication date of the headline, and (g) the pilot data obtained for the headline.

Pilot Study 2

To replace outdated headlines in our first set, we continued our headline search from April 2021 to May 2022, following the

(*Appendices continue*)

same criteria for the identification of headlines. Using the same prescreening criteria, we narrowed our data base to 105 headlines to be included our second pilot study. We then recruited 120 self-identified Republicans and 120 self-identified Democrats via Prolific and asked them the same two questions for each headline. Following the procedures in our first pilot study, we first excluded all pro-Democrat headlines with a mean partisanship score > 3.00 among either Democrats or Republicans, and all Pro-Republican headlines with a mean partisanship score < 5.00 among either Democrats or Republicans. Next, we eliminated headlines with a mode partisanship score of 4 among either Democrats or Republicans. After applying these criteria, the list of headlines identified in our second pilot study included nine true pro-Democrat headlines, 15 true pro-Republican headlines, seven false pro-Democrat headlines, and 19 false pro-Republican headlines. These headlines were used to replace outdated headlines in our first headline set by matching false pro-Republican headlines with false pro-Democrat headlines in terms of familiarity to obtain a revised list of 15 headlines per category. The revised list of headlines and the data of the second pilot study are available at https://osf.io/d2rne/.

Received September 8, 2022
Revision received January 6, 2023
Accepted January 13, 2023 ■