EUROPEAN REVIEW OF SOCIAL PSYCHOLOGY
https://doi.org/10.1080/10463283.2026.2675803

Susceptibility to misinformation as a signal detection problem

Bertram Gawronski[a], Nyx L. Ng[a], Lea S. Nahon[b], Tyler J. Hubeny[a] and Fabian M. Wurzinger[a]

[a]Department of Psychology, University of Texas at Austin, Austin, TX, USA; [b]Department of Psychology, University of Geneva, Geneva, Switzerland

ABSTRACT

Why do people fall for misinformation, and what can be done about it? The current article reviews a programme of research that has used a signal-detection framework to identify three distinct factors in truth judgements of true and false information: (a) accurate discernment of true and false information (*truth sensitivity*); (b) a general tendency to judge information as true versus false (*overall threshold*); and (c) a tendency to accept information that is congruent versus incongruent with one's views (*myside bias*). A first level of analysis addresses whether people accept false information as true due to low truth sensitivity, low overall threshold, or myside bias. A second level of analysis addresses the psychological determinants of truth sensitivity, overall threshold, and myside bias. The current article reviews key findings at each level of analysis and discusses their implications for why people fall for misinformation and what could be done about it.

ARTICLE HISTORY Received 6 August 2025; Accepted 13 May 2026

KEYWORDS Misinformation; Myside Bias; Signal Detection Theory; Truth Judgements

Misinformation is often seen as responsible for a wide range of behaviours with detrimental outcomes for individuals and society (Ecker et al., 2025; Lewandowsky et al., 2012). Examples include the presumed contribution of misinformation to vaccine hesitancy, the attack on the U.S. Capitol on 6 January 2021, the use of unproven substances to treat coronavirus infections during the COVID-19 pandemic, and the anti-immigration riots following the Southport stabbings in the United Kingdom in the summer of 2024. Concerns about these and other cases have led to a multidisciplinary effort to tackle the spread of misinformation and its impact (Kozyreva et al., 2024; Lazer et al., 2018). Two central questions in psychological research related to this endeavour are: why do people fall for misinformation, and what can be done about it? The current article reviews findings of a research

CONTACT Bertram Gawronski gawronski@utexas.edu Department of Psychology, University of Texas at Austin, 108 E Dean Keeton, Austin, TX 78712, USA

programme that has used Signal Detection Theory (SDT; Green & Swets, 1966) as a framework to address these questions. While our review focuses primarily on research conducted in our own lab, it also covers relevant studies by other researchers who used SDT in their works. The main goals of our review are to: (a) conceptually integrate the findings of extant work, (b) highlight the insights that have been gained from this work, and (c) discuss implications of these insights for basic questions regarding the psychological underpinnings of misinformation susceptibility (i.e., why do people fall for misinformation?) as well as applied questions on how to tackle misinformation susceptibility (i.e., what can be done about it?).

Signal-detection framework

Research that has used SDT to study susceptibility to misinformation focuses on instances of misinformation that can be unambiguously classified as false. While some studies have utilised SDT to study the sharing of true and false information (e.g., Gawronski et al., 2023a, 2023b), the current review focuses on beliefs in true and false information, operationalised as judgements of true and false information as true or false (see Brashier & Marsh, 2020).[1] Drawing on signal-detection terminology, a correct judgement of true information as true can be described as a *hit*; a correct judgement of false information as false can be described as a *correct rejection*; an incorrect judgement of true information as false can be described as a *miss*; and an incorrect judgement of false information as true can be described as a *false alarm* (see Table 1). Research on misinformation susceptibility is essentially concerned with false alarms: why do people judge false information as true?

According to our signal-detection framework, three factors can shape the extent to which people judge false information as true (see

Table 1. Four potential cases in truth judgements of true and false information as either true or false.

	Judged "True"	Judged "False"
True Information	HIT	MISS
False Information	FALSE ALARM	CORRECT REJECTION

[1]Other types of information that have been described as misinformation include information that is true but misleading and biased information that lacks context (Van der Linden et al., 2025). While prior misinformation research using SDT has adopted a narrower definition of misinformation as false information, it is possible to expand the use of SDT to other types of information by classifying the focal information in terms of alternative categories (e.g., information that is misleading versus not misleading) and using tasks that do not ask participants to judge the truth of the presented information (e.g., tasks that ask participants if the presented information is misleading or if they would share the presented information).

Gawronski et al., 2024). The first factor, called *truth sensitivity*, refers to how good people are at discerning true from false information (Batailler et al., 2022; Gawronski et al., 2024). A person with high truth sensitivity would correctly judge a lot of false information as false (i.e., high rate of correct rejections) and, at the same time, correctly judge a lot of true information as true (i.e., high rate of hits). Conversely, a person with low truth sensitivity would mistakenly judge a lot of false information as true (i.e., high rate of false alarms) and, at the same time, mistakenly judge a lot of true information as false (i.e., high rate of misses). Mathematically, truth sensitivity is captured by SDT's *d'* index, which is calculated as follows:

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

In this equation, *H* represents the proportion of true information judged as true; *FA* represents the proportion of false information judged as true. Both proportions are converted to follow a quantile function for a *z*-distribution, such that a proportion of 0.5 is converted to a *d'* score of zero (reflecting chance responses). Higher *d'* scores reflect greater accuracy in discerning true from false information (i.e., higher truth sensitivity). Conceptually, *d'* scores reflect the distance between the distributions for true and false information along the judgement dimension of perceived veracity (see Figure 1). Distributions that are closer together along the perceived-veracity dimension reflect a lower sensitivity, indicating that participants' ability in correctly discriminating between true and false information is relatively low (see Figure 1, upper panel). Distributions that are further apart along the perceived-veracity dimension reflect a higher sensitivity, indicating that participants' ability in correctly discriminating between true and false information is relatively high (see Figure 1, lower panel).

The second factor, called *acceptance threshold*, refers to the tendency to accept (vs. reject) information regardless of whether it is true or false (Batailler et al., 2022; Gawronski et al., 2024). A person with a high acceptance threshold would correctly judge a lot of false information as false (i.e., high rate of correct rejections) and, at the same time, mistakenly judge a lot of true information as false (i.e., high rate of misses). Conversely, a person with a low acceptance threshold would mistakenly judge a lot of false information as true (i.e., high rate of false alarms) and, at the same time, correctly judge a lot of true information as true (i.e., high rate of hits). Mathematically, acceptance threshold is captured by SDT's *c* index, which is calculated as follows:

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

Calculated in this manner, *c* scores higher than zero indicate a greater likelihood to judge information as false than to judge information as true;

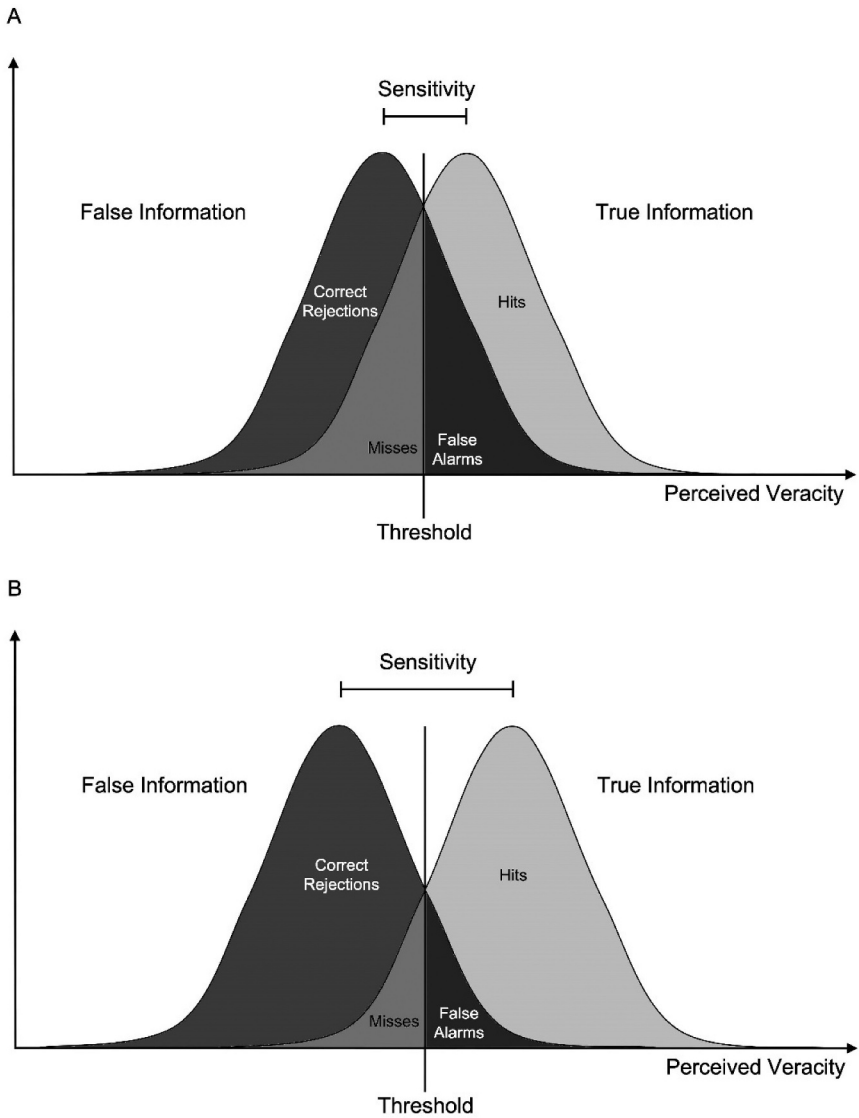

Figure 1. Graphical depiction of *truth sensitivity* within Signal Detection Theory, reflecting the distance between distributions of judgements about true versus false information along the dimension of perceived veracity. Distributions that are closer together along the perceived-veracity dimension reflect lower sensitivity, indicating that participants' ability in correctly discriminating between true and false information is relatively low (upper panel). Distributions that are further apart along the perceived-veracity dimension reflect higher sensitivity, indicating that participants' ability in correctly discriminating between true and false information is relatively high (lower panel). Figure adapted from Gawronski et al. (2025). Reprinted with permission.

c scores lower than zero indicate a greater likelihood to judge information as true than to judge information as false. A score of zero reflects an equal likelihood to judge information as true versus false. Conceptually, *c* scores reflect the threshold along the dimension of perceived veracity at which a participant decides to switch their decision (see Figure 2). When judging information as true (vs. false), *c* scores indicate 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 low threshold indicates that a participant is generally more likely to judge information as true (see Figure 2, upper panel), whereas a high threshold indicates that a participant is generally less likely to judge information as true (see Figure 2, lower panel).

The third factor, called *myside bias*, refers to the tendency to accept information that is congruent with one's personal views and to reject information that is incongruent with one's personal views (Batailler et al., 2022; Gawronski et al., 2024). This tendency is reflected in differential acceptance thresholds, in that people may show a lower acceptance threshold for information congruent with their personal views compared to information incongruent with their personal views. Mathematically, myside bias can be calculated as the difference between *c* scores for attitude-incongruent and attitude-congruent information:

$$\textit{myside bias} = c_{\text{incongruent}} - c_{\text{congruent}}$$

Calculated in this manner, scores higher than zero reflect a lower acceptance threshold for attitude-congruent than attitude-incongruent information. Scores lower than zero reflect a lower acceptance threshold for attitude-incongruent than attitude-congruent information. A score of zero reflects equal acceptance thresholds for attitude-congruent and attitude-incongruent information. Myside-bias scores conceptually align with the original definition of myside bias as people's tendency to "evaluate evidence, generate evidence, and test hypotheses in a manner biased towards their own prior opinions and attitudes" (Stanovich et al., 2013, p. 259).

In sum, an analysis drawing on SDT suggests that truth judgements of true and false information can be shaped by three factors: truth sensitivity, overall threshold, and myside bias. An important aspect of the three factors is that they are conceptually independent, in that each can vary without the other (see Gawronski, 2021; Gawronski et al., 2024). For example, differences in overall threshold do not necessarily affect truth sensitivity, because a higher overall threshold involves not only a lower rate of false alarms but also a lower rate of hits. Similarly, differential levels of myside bias do not necessarily affect overall threshold, because stronger myside bias can involve both a higher acceptance threshold for attitude-incongruent information and

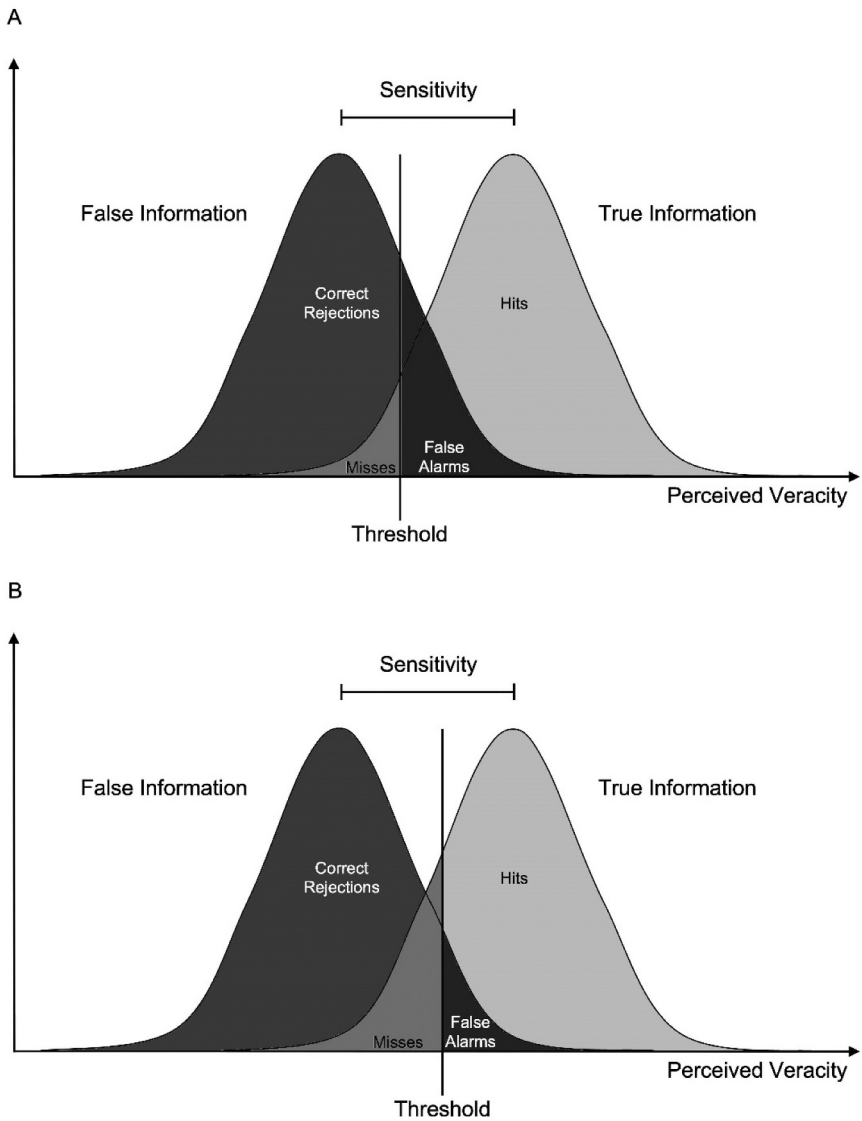

Figure 2. Graphical depiction of *acceptance threshold* within Signal Detection Theory, reflecting the threshold along the dimension of perceived veracity at which a participant decides to switch their decision. When judging truth, the threshold indicates the degree of veracity the participant must perceive 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 low threshold would indicate that a participant is generally more likely to judge information as true (upper panel), whereas a high threshold would indicate that a participant is generally less likely to judge information as true (lower panel). Figure adapted from Gawronski et al. (2025). Reprinted with permission.

a lower acceptance threshold for attitude-congruent information, leading to a compensatory effect on overall-threshold scores. Finally, differential levels of myside bias do not necessarily affect truth sensitivity, because myside bias involves (a) a higher false-alarm rate and a higher hit rate for attitude-congruent information and (b) a lower false-alarm rate and a lower hit rate for attitude-incongruent information, leading to an overall compensatory effect on truth-sensitivity scores. Thus, the three factors play distinct roles in judgements of false information as true. By providing a tool to quantify the three factors, our signal-detection framework goes beyond other approaches that exclusively focus on judgements of false information as true (i.e., false-alarm rates in our SDT framework) or accurate discernment between true and false information (i.e., truth sensitivity in our SDT framework) without considering the distinct contributions of truth sensitivity, overall threshold, and myside bias to judgements of false information as true.

Expanding on the distinction between truth sensitivity, overall threshold, and myside bias, research on misinformation susceptibility can be understood as involving two levels of analysis that differ in terms of the focal phenomena that need to be explained (i.e., *explanandum*) and the explanatory constructs proposed to explain the focal phenomena (i.e., *explanans*). At the first level of analysis (see arrows in lower part of Figure 3), the to-be-explained phenomenon is judgements of false information as true, with truth sensitivity, overall threshold, and myside bias serving as distinct (yet not mutually exclusive) explanations (see Gawronski et al., 2023a, 2024). The critical point at this level of analysis is that people may accept false

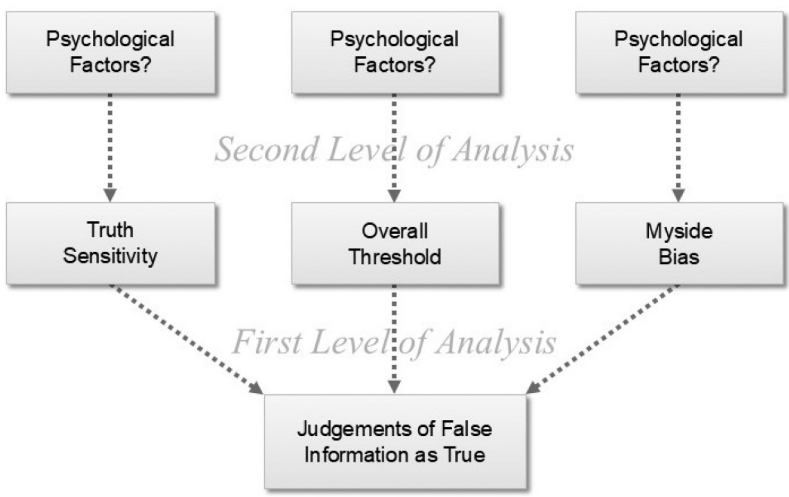

Figure 3. Two levels of analysis in understanding misinformation susceptibility from a signal-detection perspective.

information as true because they show low truth sensitivity, a low overall threshold, or myside bias, and analyses using SDT allow researchers to quantify each factor. At the second level of analysis (see arrows in upper part of Figure 3), truth sensitivity, overall threshold, and myside bias are the focal phenomena that need to be explained, with various psychological factors serving as potential explanations (see Gawronski et al., 2023a, 2024). The critical point at this level of analysis is that the three factors underlying judgements of false information as true likely have distinct psychological determinants, and analyses using SDT allow researchers to identify these determinants.

In the following sections, we first review evidence related to the first level of analysis, addressing whether people accept false information as true due to low truth sensitivity, low overall threshold, or myside bias. Expanding on this discussion, we review evidence related to the second level of analysis, addressing the psychological determinants of truth sensitivity, overall threshold, and myside bias, respectively.

First level of analysis

Expanding on the notion that judgements of false information as true can be independently shaped by truth sensitivity, overall threshold, and myside bias, a central question concerns the average levels of the three factors. How bad are people at discerning true from false information? How strong is people's general tendency to accept information? And how strong is people's tendency to accept information that is congruent with their views and to reject information that is incongruent with their views?

To address these questions, we analysed average scores of truth sensitivity, overall threshold, and myside bias in all relevant studies from our lab. To account for large variability in score distributions, we also calculated standardised mean difference scores and submitted them to an internal meta-analysis to compute average effect sizes of truth sensitivity, overall threshold, and myside bias, respectively.[2] Effect sizes were calculated as Cohen's *d*s comparing average scores on each factor to a reference point of zero.

The topics included political (mis)information and (mis)information about Covid-19 vaccines, with multiple sets of distinct stimuli in the studies on political (mis)information. The stimulus sets in the studies on (mis)information about Covid-19 vaccines were largely identical with small changes to a subset of items. The stimuli were individual headlines or

[2]The meta-analytic calculations were conducted using the R package *metafor* version 4.4.0 (Viechtbauer, 2010). To account for between-study variation in topics and stimuli, random-effects models were fitted using restricted maximum-likelihood estimation. The analysis codes and information on the availability of the data are available at https://osf.io/n5tbv/.

statements that were gathered from the internet and thoroughly fact-checked (for more details on the selection of stimuli in studies on political misinformation, see Appendix of Gawronski et al., 2023a; for more details on the selection of stimuli in studies on misinformation about Covid-19 vaccines, see Appendix A of; Nahon et al., 2024). Except for one study that was run on CloudResearch (Gawronski et al., 2023b, Study S1), all samples were recruited on Prolific Academic. All studies on political (mis)information included participants from the United States (Gawronski et al., 2023a, 2023b; Hubeny, Nahon, Ng, et al., 2026; Nahon et al., 2026; Ng et al., 2026). Several studies on (mis)information about Covid-19 vaccines included participants from the United States and the United Kingdom (Hubeny, Nahon, Ng, et al., 2026; Nahon et al., 2024); some included only participants from the United States (Nahon et al., 2026). In studies on political (mis)information, myside bias was operationalised via participants' self-reported political affiliation as Democrat or Republican and classifications of the stimuli as either pro-Democrat or pro-Republican based on pilot tests. In studies on (mis)information about Covid-19 vaccines, myside bias was operationalised via participants' self-reported attitudes towards Covid-19 vaccines (i.e., favourable vs. unfavourable) and classifications of the stimuli as either pro-vaccine or anti-vaccine based on judgements by the research team.

Although truth-sensitivity scores showed large variability, average levels of truth sensitivity in discerning true from false information were relatively high overall (see Figure 4). The meta-analytic effect across studies qualifies as large with a Cohen's *d* of 1.31. This finding aligns with the results of broader meta-analyses that used SDT to reanalyse individual participant data (IPD), showing that people can distinguish between true and false information with a remarkably high degree of accuracy (Pfänder & Altay, 2025; Sultan et al., 2024; see also Gawronski et al., 2025).[3]

Interestingly, our analysis revealed no evidence for a general tendency to accept information as true. Instead, average levels of overall-threshold scores indicate a general tendency to reject information as false (see Figure 5). While there was considerable variability in overall-threshold scores, the meta-analytic effect across studies qualifies as medium with a Cohen's *d* of 0.50. Broader IPD meta-analyses using SDT have obtained mixed evidence on this point, with one meta-analysis reporting no significant tendency in either direction (Sultan et al., 2024) and another replicating the significant tendency to reject information as false found in our studies (Pfänder & Altay, 2025). Despite these differences, the available evidence conflicts with the idea that participants in the reviewed studies fell for misinformation because of

[3]The results of Pfänder and Altay's (2025) IPD meta-analysis using SDT are reported in Appendix H of the Supplemental Materials to their article.

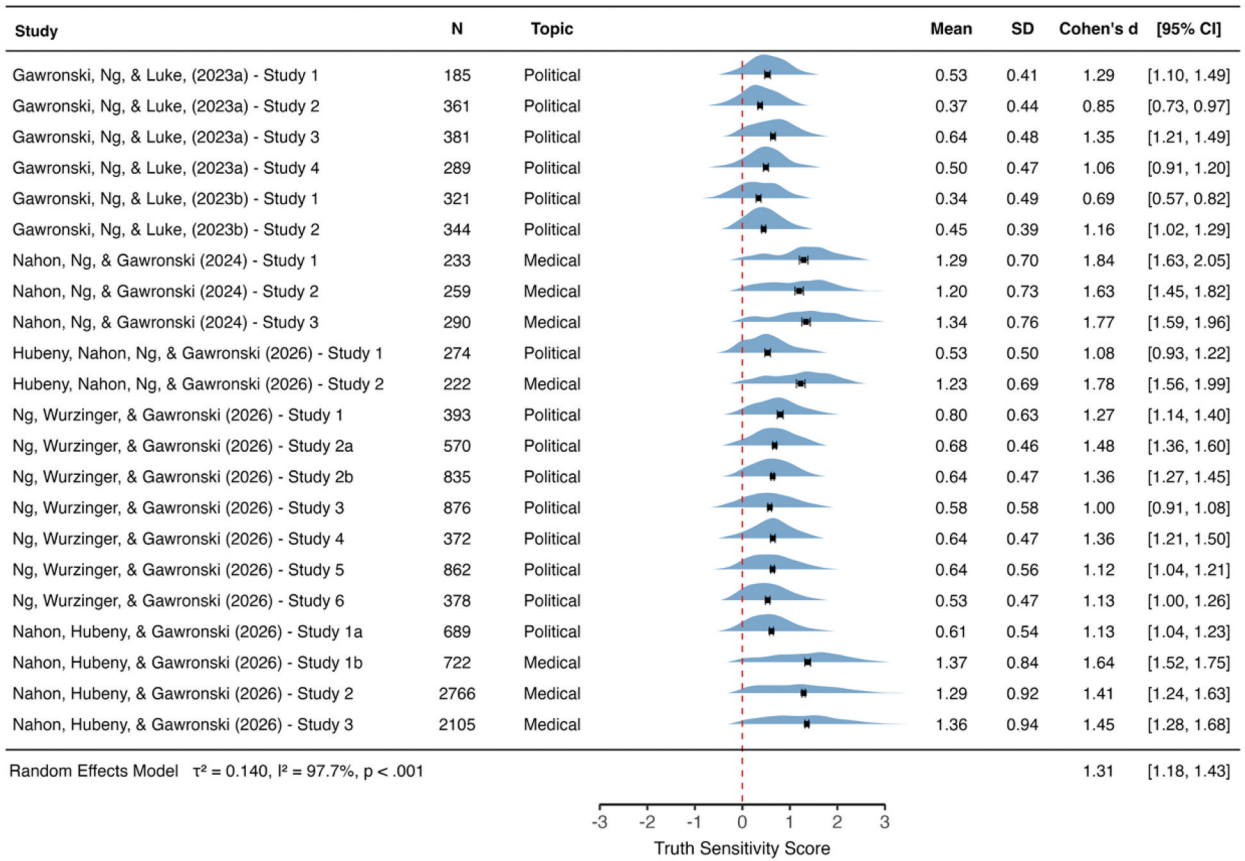

Figure 4. Distributions of truth-sensitivity scores and standardised mean differences (Cohen's *d*) across studies. The reference point at zero reflects chance-level performance in distinguishing between true and false information. Higher scores reflect greater accuracy in discerning true from false information.

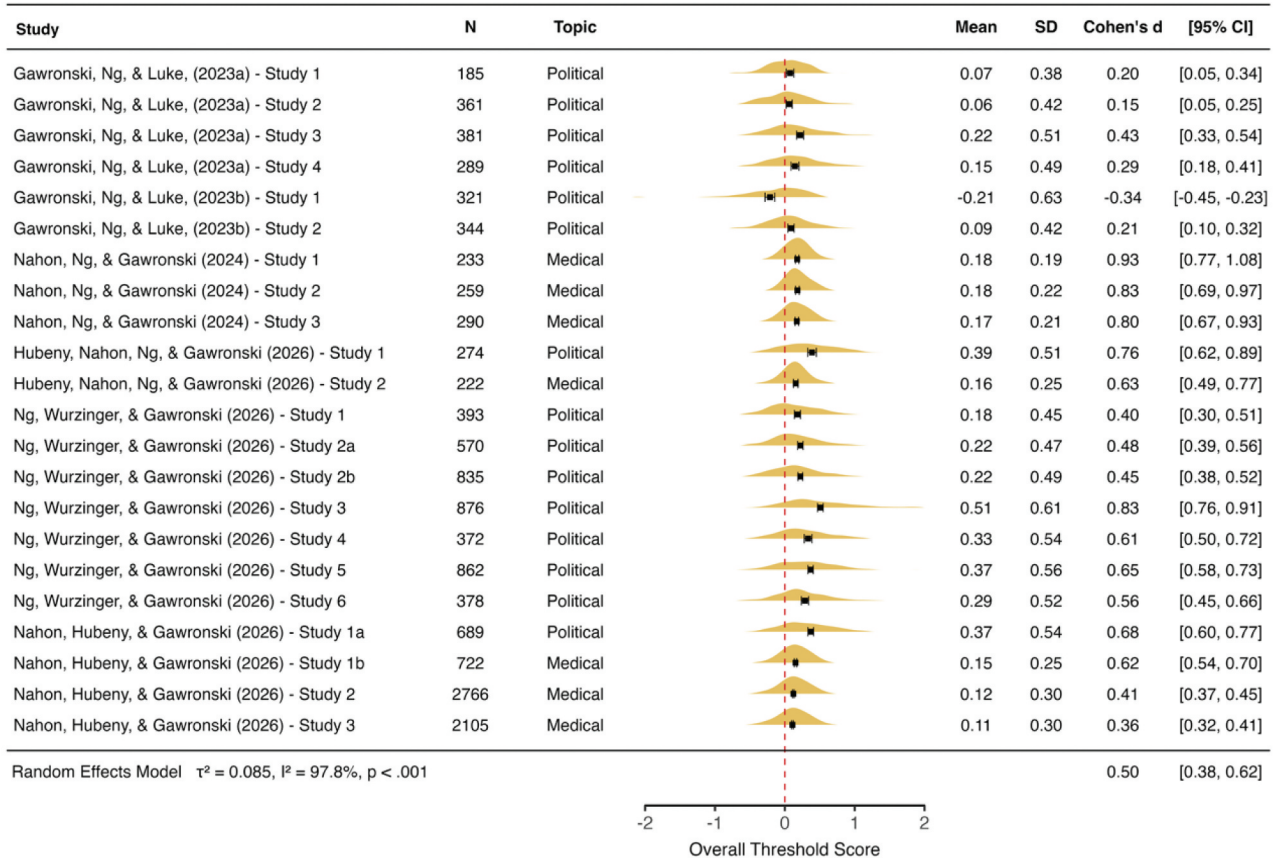

Figure 5. Distributions of overall-threshold scores and standardised mean differences (Cohen's *d*) across studies. The reference point at zero reflects an equal likelihood of judging information as true or false. Scores higher than zero reflect a tendency to judge information as false. Scores lower than zero reflect a general tendency to judge information as true.

a general tendency to accept information as true (see Brashier & Marsh, 2020; Levine, 2014). If that were the case, overall-threshold scores should be significantly lower than zero, which is not the case in our studies and in broader IPD meta-analyses using SDT.

Across all studies, we found reliable evidence for myside bias in judgements of true and false information (see Figure 6). This finding indicates that, on average, participants showed a lower acceptance threshold for attitude-congruent than attitude-incongruent information. Although there was considerable variability in myside-bias scores, the meta-analytic effect across studies qualifies as large with a Cohen's *d* of 0.95. Similar findings are reported in a broader IPD meta-analysis using SDT (Sultan et al., 2024).[4]

Together, these findings suggest that, on average, (a) people are quite good at discerning true from false information, (b) people show a general tendency to reject information as false, and (c) myside bias in judgements of true and false information is pervasive and large. From this perspective, susceptibility to misinformation seems to be primarily a product of myside bias rather than low truth sensitivity or low overall threshold (see Gawronski et al., 2025). Yet, the link between myside bias and susceptibility to misinformation is more complex, in that myside bias involves stronger susceptibility to attitude-congruent misinformation and weaker susceptibility to attitude-incongruent misinformation. In other words, the misinformation that people mistakenly accept as true is very likely to be attitude-congruent and very unlikely to be attitude-incongruent.

Second level of analysis

Expanding on the argument that judgements of false information as true can be jointly shaped by truth sensitivity, overall threshold, and myside bias (first level of analysis), an important follow-up question concerns the psychological determinants of the three factors (second level of analysis). In the following sections, we review evidence from correlational and experimental studies that address this question.

Individual-difference correlates

While average levels of truth sensitivity, overall threshold, and myside bias were all relatively high in our studies, the score distributions for each factor suggest considerable variability across participants (see Figures 4–6). The large interindividual variability raises the question of whether people with certain individual-difference characteristics are more susceptible to judging

[4]Although Pfänder and Altay (2025) do not report SDT analyses of myside bias, their meta-analysis revealed a conceptually corresponding effect on raw scores of truth judgements.

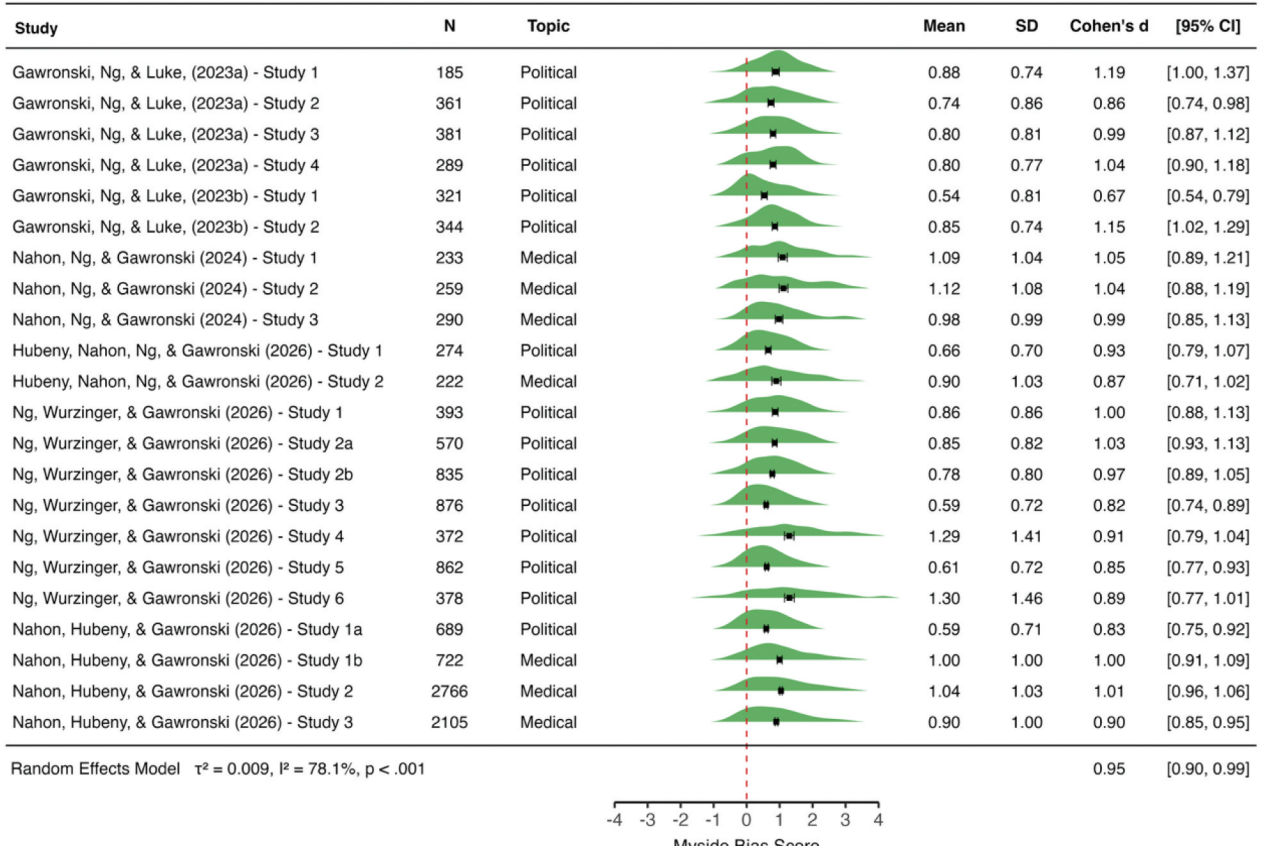

Figure 6. Distributions of myside-bias scores and standardised mean differences (Cohen's *d*) across studies. The reference point at zero reflects equal acceptance thresholds for attitude-congruent and attitude-incongruent information. Scores higher than zero reflect lower thresholds for attitude-congruent than attitude-incongruent information. Scores lower than zero reflect lower thresholds for attitude-incongruent than attitude-congruent information.

false information as true, and whether any such differences are driven by individual differences in truth sensitivity, overall threshold, or myside bias. Two studies by Hubeny, Nahon, Ng, et al. (2026) addressed this question. Participants completed a battery of 15 individual-difference measures capturing the Big Five personality traits (Soto & John, 2017), need to belong (Leary et al., 2013), cognitive reflection (Frederick, 2005; Thomson & Oppenheimer, 2016), receptivity to pseudoprofound bullshit (Pennycook et al., 2015), conspiracy mentality (Bruder et al., 2013), self-esteem (Rosenberg, 1965), grandiose narcissism (Ames et al., 2006), intellectual humility (Leary et al., 2017), actively open-minded thinking (Stanovich & Toplak, 2023), need to evaluate (Jarvis & Petty, 1996), and identification with likeminded people (McFarland et al., 2012). After completing the individual-difference measures, participants judged the truth of news headlines about political issues (Study 1) or statements about Covid-19 vaccines (Study 2). Although truth sensitivity, overall threshold, and myside bias all showed acceptable estimates of internal consistency (Cronbach's αs between .68 and 91) and varied considerably across participants, only truth sensitivity showed reliable associations with the measured individual-difference variables. Across both content domains, truth sensitivity was greater among (a) participants high in cognitive reflection, (b) participants high in actively open-minded thinking, (c) participants low in receptivity to pseudoprofound bullshit, and (d) participants low in conspiracy mentality (see also Barajas & John, 2023). None of the other 11 individual-difference dimensions showed reliable associations that replicated across content domains. Further analyses using bifactor modelling (Rodriguez et al., 2016) suggest that the obtained associations with truth sensitivity are driven by a single underlying construct, which Hubeny, Nahon, Ng, et al. (2026) called *reflective open-mindedness*, with reference to a construct originally proposed by Pennycook and Rand (2020). Together, these findings suggest that people high in reflective open-mindedness are less likely to fall for misinformation because they are better at discerning true from false information. Moreover, while there is considerable interindividual variability in overall threshold and myside bias, the individual-difference dimensions associated with these two factors are still unclear.[5]

Cognitive elaboration

One potential interpretation of the obtained association between truth sensitivity and reflective open-mindedness is that greater cognitive

[5]A recent study by Ramos and Van Boven (2025) suggests that older people show greater levels of myside bias than younger people. However, the psychological underpinnings of the obtained association remain unclear. There were no significant associations between myside bias and age in Hubeny, Nahon, Ng, et al.'s (2026) studies.

elaboration during the processing of true and false information reduces susceptibility to misinformation by increasing people's ability to distinguish between true and false information (Pennycook, 2023). Yet, an alternative interpretation is that people high in reflective open-mindedness are simply more knowledgeable than people low in reflective open-mindedness. Such knowledge differences could arise from different information-consumption habits, including differences in the overall amount of information people seek out in their daily lives or differences in the reliability of the media outlets from which they obtain information (see Grant et al., 2024). While it is difficult to rule out knowledge differences in studies using correlational designs, several experimental studies provide evidence for a causal effect of cognitive elaboration during the processing of true and false information. Specifically, these studies found that truth sensitivity was higher when participants had unlimited time to judge the truth of true and false information than when participants had to make their judgements under time pressure (Gawronski et al., 2023a; Nahon et al., 2024; Sultan et al., 2022). Time pressure had no significant effects on overall threshold and myside bias in any of these studies. The latter finding seems notable because it suggests that, while greater cognitive elaboration is effective in increasing truth sensitivity, it is ineffective in reducing myside bias (see also Batailler et al., 2022; Ludwig & Sommer, 2024; Perez Santangelo & Solovey, 2023). In fact, meta-analytic findings by Sultan et al. (2024) suggest the opposite, in that myside bias in judgements of true and false information was positively (rather than negatively) associated with cognitive reflection (see also Lois et al., 2026). This finding is consistent with the so-called *motivated-system-2 hypothesis*, which suggests that people use their cognitive capabilities to support and protect their beliefs (see Kahan, 2013). However, it is worth noting that almost all studies in Sultan et al.'s (2024) meta-analysis used correlational rather than experimental designs to investigate links between cognitive reflection and susceptibility to misinformation. Hence, it is possible that the positive association between myside bias and cognitive reflection is driven by other factors that are not directly related to cognitive elaboration during the encoding of true and false information (e.g., belief confidence; see Gawronski et al., 2023a; Nahon et al., 2024). Yet, despite this ambiguity, the available evidence provides strong support for the conclusions that (a) cognitive elaboration increases truth sensitivity, (b) cognitive elaboration does not affect overall threshold, and (c) cognitive elaboration does not reduce myside bias.

Self-affirmation

The finding that cognitive elaboration is ineffective in reducing myside bias highlights the significance of understanding the psychological underpinnings of myside bias. A common assumption is that myside bias is the product of motivated reasoning (see Kunda, 1990), in that people are motivated to support and protect beliefs that are important to them (e.g., Van Bavel & Pereria, 2018). This hypothesis is based on the idea that having cherished beliefs affirmed makes people feel good about themselves, whereas having cherished beliefs threatened makes people feel bad about themselves (Sherman & Cohen, 2006). Thus, to regulate their self-feelings, people tend to accept information that is congruent with their personal views and reject information that is incongruent with their personal views (see also Van Bavel et al., 2024).

An alternative hypothesis posits that myside bias arises from purely cognitive processes that follow the principles of Bayesian inference (Pennycook & Rand, 2021). According to this view, differences in the acceptance of attitude-congruent and attitude-incongruent information arise from their mere consistency with prior beliefs and the confidence with which these beliefs are held (i.e., Bayesian priors). If new information is consistent with strongly held beliefs, it is deemed "rational" to accept this information as true. Conversely, if new information is inconsistent with strongly held beliefs, it is deemed "rational" to reject this information as false. Because people with different identities (e.g., Republicans vs. Democrats in the United States) differ in terms of both their motivations and their prior beliefs, it is notoriously difficult to isolate the respective contributions of the two mechanisms to myside bias in judgements of true and false information (see Ditto et al., 2025; Druckman & McGrath, 2019; Tappin et al., 2020).

To disentangle the two mechanisms, Gawronski et al. (2023a) investigated the effects of self-affirmation on myside bias in truth judgements of true and false political information. The idea underlying this study was that self-affirmation should show opposite effects on myside bias depending on whether myside bias arises from motivational versus cognitive processes. From a motivational view, self-affirmation can be assumed to enhance positive feelings about the self, which should reduce the need to regulate one's self-feelings by accepting attitude-congruent information and rejecting attitude-incongruent information. From a cognitive view, on the other hand, self-affirmation can be assumed to enhance confidence in one's beliefs, which should increase the tendency to accept attitude-congruent information and to reject attitude-incongruent information. In other words, whereas a motivational account suggests that self-affirmation should decrease myside bias (via increased positive feelings about the self), a cognitive account

suggests that self-affirmation should increase myside bias (via increased confidence in one's beliefs).

Consistent with the proximal effects proposed by the two accounts, Gawronski et al. (2023a) found that self-affirmation (vs. self-threat) increased both positive feelings about the self and confidence in one's beliefs.[6] However, self-affirmation had no significant effect on myside bias. There were also no significant effects on truth sensitivity and overall threshold. The unexpected null effect on myside bias raises the interesting possibility that the two mechanisms may jointly contribute to myside bias. In this case, self-affirmation should be ineffective in either reducing or increasing myside bias, because the downstream effect of one mechanism may counteract the downstream effect of the other mechanism. To test this possibility, Gawronski et al. (2023a) conducted a multiple-mediator path analysis testing indirect effects of self-affirmation on myside bias via positive feelings about the self and confidence in one's beliefs (see Figure 7). Consistent with a cognitive account of myside bias, self-affirmation significantly increased confidence in one's beliefs, which in turn showed a significant positive association with myside bias. Evidence for

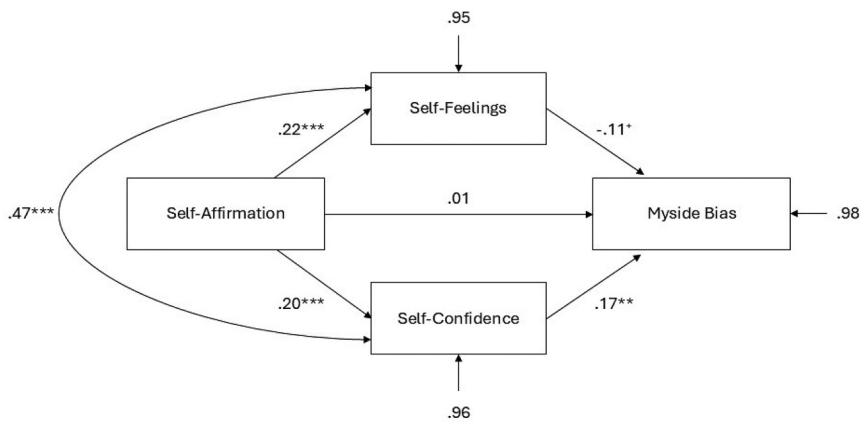

Figure 7. Results of path-model analysis testing effects self-affirmation (vs. self-threat) on myside bias in truth judgements of true and false political information via positive self-feelings and levels of self-confidence. Figure adapted from Gawronski et al. (2023a). Reprinted with permission.

[6]Prior to running the reviewed study, Gawronski et al. 2023b conducted two conceptually similar studies on effects of self-affirmation, the main difference being that these studies used a neutral control condition instead of a self-threat condition. While Gawronski et al.'s (2023a) study comparing self-affirmation to self-threat obtained a significant effect on both self-feelings and self-confidence, the two studies comparing self-affirmation to a neutral control condition found no reliable effects on the two proximal outcomes. The two studies also showed no significant effect on myside bias.

a motivational account of myside bias was mixed, in that self-affirmation significantly increased positive feelings about the self; yet the predicted negative association between positive self-feelings and myside bias was only marginal. Moreover, whereas the indirect path via self-confidence was statistically significant, the indirect path via self-feelings was only marginal. A correlational study by Nahon et al. (2024) revealed a similar pattern for (mis)information about Covid-19 vaccines. In this study, myside bias showed a significant positive association with confidence in one's beliefs, but myside bias showed no significant association with positive feelings about the self. Together, these findings support the idea that myside bias arises from cognitive processes that follow the principles of Bayesian inference. However, the findings provide no compelling evidence for the idea that myside bias arises from self-protective motivated reasoning.

Fallibility salience

While the positive association between myside bias and confidence in one's beliefs is consistent with a cognitive account of myside bias, the relevant findings reviewed thus far are merely correlational, raising the question of whether confidence increases myside bias or vice versa. On the one hand, it is possible that greater confidence leads to greater myside bias. On the other hand, greater myside bias could also lead to greater confidence. A closely related question is whether interventions targeting confidence would have the potential to reduce myside bias in truth judgements of true and false information. To address these questions, we turned to the literature on intellectual humility (for a review, see Porter et al., 2022). While there are disagreements about the most appropriate way to conceptualise and measure intellectual humility, there is consensus about the centrality of acknowledging the limits of one's knowledge (see Porter, 2025), which can be understood as the direct opposite of high confidence in one's beliefs. An effective procedure to increase acknowledgement of one's limited knowledge is to increase the salience of one's fallibility, for example by giving participants multiple-choice questions with a seemingly obvious, yet incorrect response option and a non-obvious correct response option (Koetke et al., 2023). The idea underlying this procedure is that participants are highly confident in selecting the seemingly obvious, yet incorrect option, and then become aware of their limited knowledge when they learn about the non-obvious correct answer.

In a series of four studies, Nahon et al. (2026) used this procedure to test (a) the presumed causal effect of confidence on myside bias and (b) the potential of fallibility-salience interventions for reducing myside bias in truth judgements of true and false information. To this end, Nahon et al. presented

participants with a series of five multiple-choice questions with four response options. In one condition, the four response options included a seemingly obvious, yet incorrect answer, with the correct answer being non-obvious (high fallibility-salience condition). In another condition, the four response options were designed so that the correct answer is obvious (low fallibility-salience condition). After completing the multi-choice task, participants received feedback on their performance and detailed information about the correct answer to each question, followed by measures capturing acknowledgement of limited knowledge and general confidence in one's beliefs. Next, participants completed a truth judgement task in which they judged the truth of true and false statements that were congruent or incongruent with their self-reported political affiliation (i.e., Democrat vs. Republican) or their self-reported attitudes towards Covid-19 vaccines (i.e., positive vs. negative). In two of the four studies, the multiple-choice questions and the statements in the truth-judgement task involved the same content domain (i.e., medical issues). The remaining two studies manipulated whether the content domain of the multiple-choice questions matched or mismatched the content domain of the statements in the truth judgement task (i.e., medical – medical, political – medical, medical – political, political – political).

Supporting the effectiveness of the manipulation, participants in the high (vs. low) fallibility-salience condition showed substantially lower accuracy in answering the multiple-choice questions and significantly greater acknowledgement of limited knowledge in the content domain of the multiple-choice questions. This finding replicated across all four studies. An internal meta-analysis of the data from the four studies also provided evidence for generalisation, indicating that participants in the high (vs. low) fallibility-salience condition showed weaker general confidence in their beliefs and greater acknowledgement of limited knowledge in a content domain that did not match the content domain of the multiple-choice questions. Regarding our main question, fallibility salience significantly reduced myside bias in three of the four studies. An internal meta-analysis of the data from all four studies revealed a small but statistically significant effect of fallibility salience on myside bias, indicating that myside bias was significantly weaker in the high (vs. low) fallibility-salience condition. Interestingly, while general confidence showed a significant positive association with myside bias in all four studies (replicating earlier findings by Gawronski et al. 2023a; Nahon et al., 2024), there was no reliable meta-analytic association between myside bias and acknowledgement of limited knowledge in the matching content domain. There were no reliable effects of fallibility salience on truth sensitivity and overall threshold (see also Lyons et al., 2025).

Together, these findings suggest that: (a) fallibility salience is effective in increasing acknowledgement of limited knowledge in matching content domains; (b) fallibility salience reduces myside bias, but the effect is relatively

small; (c) acknowledgement of limited knowledge in matching content domains is unrelated to myside bias, and therefore the observed increase in acknowledgement of limited knowledge in matching content domains does not explain the observed reduction in myside bias; (d) fallibility salience reduces general confidence; (e) general confidence is positively associated with myside bias; and hence (f) fallibility salience likely reduces myside bias by reducing general confidence, not by increasing acknowledgement of limited knowledge in matching content domains. These conclusions support the ideas that belief confidence increases myside bias in truth judgements of true and false information, and that interventions to reduce confidence may have the potential to reduce myside bias.

Group membership

The findings regarding belief confidence and fallibility salience are consistent with accounts that attribute myside bias in judgements of true and false information to cognitive processes that follow the principles of Bayesian inference. We found little to no evidence in these studies for accounts that attribute myside bias to identity-protective motivated reasoning. Yet, it seems premature to dismiss identity-protective motivated reasoning as a potential factor underlying myside bias, because lack of positive evidence is not the same as negative evidence (see Gawronski & Bodenhausen, 2015). Moreover, in the reviewed studies, the predictions derived from motivational accounts were all based on auxiliary assumptions about the role of positive self-feelings. Thus, the lack of positive evidence in these studies could also be due to faulty auxiliary assumptions rather than faulty core assumptions of the motivated-reasoning account (see Gawronski & Bodenhausen, 2015).

To tackle this issue, Hubeny, Nahon, and Gawronski (2026) conducted two studies that tested effects of randomly assigned group identities on myside bias in truth judgements of true and false information. A major ambiguity in misinformation research using pre-existing group identities is that such identities involve differences in both preferences and knowledge (see Ditto et al., 2025; Druckman & McGrath, 2019; Tappin et al., 2020). For example, American participants who identify as a Democrat or Republican differ not only in terms of the political information they prefer (e.g., Democrats prefer information with a pro-Democrat slant, whereas Republicans prefer information with a pro-Republican slant), but also in terms of the information ecologies they live in (e.g., Democrats tend to receive more information from Democrat-leaning news sources, whereas Republicans tend to receive more information from Republican-leaning news sources), which creates knowledge differences that favour the political ingroup over the political outgroup. Hence, myside bias among Democrats and Republicans (and other pre-existing groups) may be due to either

motivated reasoning or differential knowledge (see Ditto et al., 2025; Druckman & McGrath, 2019; Tappin et al., 2020). By using randomly assigned identities, Hubeny, Nahon, and Gawronski (2026) isolated the effects of identity-protective motivated reasoning by eliminating pre-existing knowledge differences.

To this end, American participants received information about conflicts between two European countries. In Experiment 1, the two countries were France and the United Kingdom. In Experiment 2, the two countries were Greece and Spain. Participants then completed a personality test and were randomly assigned to one of two teams representing the two countries based on bogus feedback about their scores on the personality test. Participants in a control condition underwent the same procedure without being assigned to a team. Next, all participants judged the truth of true and false statements about the two countries, half of which were congenial to one country while the other half were congenial to the other country. As predicted, randomly assigned group identities produced a pattern of myside bias, in that acceptance thresholds were significantly lower for statements with a pro-ingroup slant compared to statements with a pro-outgroup slant. Acceptance thresholds in the control condition did not significantly differ depending on whether the statements favoured one country over the other. Truth sensitivity was unaffected by randomly assigned group identities. Together with the reviewed studies on belief confidence and fallibility salience, these results suggest that myside bias can arise from both (a) cognitive processes following the principles of Bayesian inference and (b) identity-protective motivated reasoning.

Source characteristics

The studies reviewed thus far all utilised true and false statements as target stimuli without any information about their source. This approach is different from a common practice in this area to present true and false statements together with information about their original source (e.g., in the format of Facebook posts with original images, logos, and text indicating the source). While the latter approach arguably has greater ecological validity, a major downside is that it confounds the truth status of the statement with relevant characteristics of the source (e.g., true statements being presented with a reliable source and false statements being presented with an unreliable source). Although this confound reflects a naturally occurring association (i.e., unreliable sources are more likely to convey false information than reliable sources), it is problematic for basic research on misinformation susceptibility, because it can lead to inaccurate conclusions about how source characteristics influence judgements of true and false information.

An illustrative example is a meta-analytic finding by Sultan et al. (2024), suggesting that truth sensitivity is greater when source information is present than when source information is absent. Yet, because all studies in the database for Sultan et al.'s meta-analysis included the abovementioned confound between truth status and source characteristics, it is possible that the reported effect is driven by a lower acceptance threshold for information from sources that are perceived to be reliable (vs. unreliable) rather than a genuine effect on truth sensitivity *per se*. Disentangling the two possibilities requires fully crossed designs in which both true and false information is presented with sources that are perceived to be either reliable or unreliable. In such designs, participants may be more likely to judge information as true when it comes from a reliable source, regardless of whether the information is true or false. Conversely, participants may be more likely to judge information as false when it comes from an unreliable source, regardless of whether the information is true or false. Needless to say, such an effect on acceptance thresholds would be quite different from the presumed effect on truth sensitivity suggested by Sultan et al.'s (2024) meta-analysis.

A series of seven experiments by Ng et al. (2026) provide deeper insights into how source characteristics affect truth judgements of true and false political information. In a first experiment, Ng et al. aimed to replicate the pattern obtained in Sultan et al.'s meta-analysis, comparing a condition in which truth status is confounded with perceived source reliability to a condition in which no source information was provided. Consistent with the pattern in Sultan et al.'s meta-analysis, truth sensitivity was significantly greater when truth status was confounded with perceived source reliability than when no source information was provided. Four follow-up experiments tested effects of source reliability using fully crossed designs that do not confound truth status with source characteristics. In these studies, perceived source reliability affected truth judgements via acceptance thresholds rather truth sensitivity *per se*, in that participants showed lower acceptance thresholds for information from sources perceived to be reliable (vs. unreliable). Yet, source effects were limited to conditions in which participants saw information from more than one source, suggesting that source effects depend on conditions that promote comparisons of sources (see Mussweiler, 2003). Interestingly, perceived source reliability had no effect on myside bias, in that participants showed lower acceptance thresholds for information that was congruent (vs. incongruent) with their self-reported political identity (i.e., Democrat vs. Republican) regardless of whether this information was presented with a source perceived to be reliable or unreliable. In other words, participants readily accepted attitude-congruent information even when it came from a source perceived to be unreliable, and they readily rejected attitude-incongruent information even when it came from a source perceived to be reliable. When perceived source reliability was not confounded with truth status, perceived source reliability showed no reliable effects on truth sensitivity.

To isolate effects of perceived source reliability, the described experiments by Ng et al. utilised sources that participants perceived as politically neutral, yet different in the perceived likelihood of reporting false information (i.e., National Enquirer vs. Forbes). Expanding on the findings of these experiments, Ng et al. conducted two follow-up studies to investigate effects of perceived source partisanship. To this end, American participants who identified as either a Democrat or a Republican were presented with true and false information that had either a pro-Democrat or a pro-Republican slant, and this information was presented with either a Democrat-leaning or a Republican-leaning source (i.e., CNN vs. Fox News). Based on classic attribution theory (Kelley & Michela, 1980), Ng et al. reasoned that effects of perceived source partisanship may differ from effects of perceived source reliability, in that effects of source partisanship depend on the political slant of the focal information. Conceptually, source partisanship involves (a) a low probability of attitude-congruent information being reported by an attitude-incongruent source and (b) a low probability of attitude-incongruent information being reported by an attitude-congruent source. Thus, in both cases, recipients may assume that the focal information must be true – otherwise, the source would not report it. In attributional terms, inconsistency between the political slant of information and the political leaning of its source may augment the diagnostic value of the information, whereas consistency between the two suggests that its diagnostic value should be discounted (Kelley & Michela, 1980). From this perspective, acceptance thresholds for attitude-congruent information should be lower for attitude-incongruent than attitude-congruent sources. Conversely, acceptance thresholds for attitude-incongruent information should be lower for attitude-congruent than attitude-incongruent sources.

Ng et al.'s two experiments on effects of perceived source partisanship did not confirm these predictions. Instead, perceived source partisanship showed effects that were functionally identical to the ones obtained for perceived source reliability. Specifically, participants showed a lower acceptance threshold for information from attitude-congruent (vs. attitude-incongruent) sources irrespective of the political slant of the focal information. Like the effects of perceived source reliability, effects of perceived source partisanship were limited to conditions in which participants saw information from more than one source, suggesting that source effects depend on conditions that promote comparisons of sources (see Mussweiler, 2003). Finally, perceived source partisanship had no effect on myside bias, in that participants showed a lower acceptance threshold for information that was congruent (vs. incongruent) with their self-reported political identity regardless of whether this information was presented with a source perceived to be attitude-congruent or attitude-incongruent. Perceived source partisanship showed no reliable effects on truth sensitivity.

Prior exposure

Research on the illusory-truth effect suggests that prior exposure to a statement increases its perceived veracity (for a review, see Unkelbach et al., 2019). A common explanation for this effect is that prior exposure increases processing fluency, which in turn increases perceived veracity (Schwarz et al., 2007). To investigate the potential contribution of prior exposure to misinformation susceptibility, Pennycook et al. (2018) presented participants with true and false news headlines and asked them if they would share the stories online. Afterwards, participants completed a truth-judgement task that included the headlines from the prior task as well as novel headlines that had not yet been presented. Consistent with prior research on the illusory-truth effect, participants were more likely to judge false headlines as true when participants had been exposed to the headlines before than when they had not been previously exposed to the headlines.

A reanalysis of Pennycook et al.'s (2018) data using SDT suggests that prior exposure influences truth judgements via acceptance threshold, in that participants' general tendency to reject all information as false was lower for headlines that had been presented before than for headlines that had not been previously presented (Batailler et al., 2022). Interestingly, an integrative data analysis (Curran & Hussong, 2009) including all studies by Pennycook et al. (2018) also revealed a negative effect of prior exposure on truth sensitivity, in that accurate discernment between true and false headlines was lower for headlines that had been presented before than for headlines that had not been presented before (Batailler et al., 2022). However, this effect was relatively weak and not reliable across individual studies. Whether and in what direction prior exposure may affect myside bias remains an open question given that Pennycook et al. (2018) did not manipulate attitude-congruence of the headlines. Yet, there is evidence that prior exposure can increase perceived veracity even for information that contradicts prior knowledge (Udry & Barber, 2024) and attitudes (Jiang et al., 2024). These findings suggest that prior exposure may decrease acceptance threshold not only for attitude-congruent information but also for attitude-incongruent information, thereby decreasing overall threshold without affecting myside bias. Future research may address this question.

Psychological inoculation

Based on concerns about potentially harmful effects of misinformation (see Ecker et al., 2025; Lewandowsky et al., 2012), a large body of research has investigated the effectiveness of various interventions to reduce susceptibility to misinformation (for reviews, see Ecker et al., 2022; Kozyreva et al., 2024). A prominent example is psychological inoculation, which is based on the idea that resistance to strong doses of misinformation could be achieved via prior administration of

a weak dose (Van der Linden, 2024). To broaden the reach of inoculation interventions, researchers have developed gamified versions in which game players are exposed to weak examples of misinformation in computerised games. A general finding in this line of work is that gamified inoculation interventions effectively reduce susceptibility to misinformation (e.g., Basol et al., 2020; Maertens et al., 2021; Roozenbeek & van der Linden, 2019; Roozenbeek et al., 2020).

While this finding has inspired broad applications of gamified inoculation interventions to combat susceptibility to misinformation (for a review, see Van der Linden, 2024), a reanalysis of extant data using SDT has raised questions about their effectiveness. Specifically, Modirrousta-Galian and Higham (2023) found that gamified inoculation interventions merely increase overall threshold without increasing truth sensitivity (see also Hoes et al., 2024). This pattern was found in every dataset that was available at the time of Modirrousta-Galian and Higham's reanalysis except for one dataset by Iyengar et al. (2023). The pattern also emerged in a meta-analysis of the data from all studies available at that time. Expanding on the results of Modirrousta-Galian and Higham (2023) reanalysis, Seabrooke et al. (2026) conducted a close replication of Iyengar et al.'s (2023) study that additionally controlled for a design confound pertaining to the use of stimuli in the original study's pre-test and post-test. Seabrooke et al. found no significant intervention effect on truth sensitivity regardless of the design confound. Moreover, in the conditions that matched the confounded design of Iyengar et al.'s original study, the intervention significantly increased overall threshold, but this effect was not significant in the fully crossed design that did not include the design confound of Iyengar et al.'s (2023) original study.

More recent IPD meta-analyses that used SDT to analyse data from larger sets of studies have found mixed evidence on how inoculation affects susceptibility to misinformation. While one (yet unpublished) meta-analysis replicated Modirrousta-Galian and Higham's (2023) finding that inoculation increases overall threshold without increasing truth sensitivity (Sun et al., 2025), another meta-analysis found that inoculation increases truth sensitivity without affecting overall threshold (Simchon et al., 2026). Because the two meta-analyses differ in terms of the employed data-inclusion criteria and various data-analytic choices, it is difficult to determine what produced these diametrically opposing results. Moreover, both meta-analyses are prone to criticism for using somewhat unusual data-inclusion criteria.[7] Greater clarity about the effects of inoculation could be achieved with a more comprehensive meta-analysis using a multiverse approach. To the extent that multi-verse analyses produce the same outcome with different data-analytic approaches, stronger conclusions could

[7]While Simchon et al.'s (2026) meta-analysis could be criticised for including published and unpublished data only from the authors' extended research group, Sun et al.'s (2025) meta-analysis could be criticised for including only published (but not unpublished) data.

be drawn about whether inoculation increases truth sensitivity or overall threshold (or both). Future studies manipulating the attitude-congruence of the to-be-judged information would also be helpful to determine whether (and, if so, in what direction) inoculation influences myside bias (see Loughnan et al., 2026).

Summary

In sum, the reviewed findings suggest that reflective open-mindedness at the individual level and cognitive elaboration during the processing of true and false information are associated with greater truth sensitivity. Yet, neither reflective open-mindedness nor cognitive elaboration affect overall threshold and myside bias. If anything, myside bias increases (rather than decreases) as a function of cognitive elaboration. The available evidence further suggests that myside bias can arise from cognitive processes that follow principles of Bayesian inference, in that greater confidence in one's beliefs (i.e., strong Bayesian priors) leads to greater myside bias. In addition, myside bias can arise from identity-protective motivated reasoning, in that randomly assigned identities can produce myside bias in the absence of identity-related differences in prior knowledge. Source characteristics such as perceived reliability and perceived partisanship affect truth judgements via overall threshold rather than truth sensitivity *per se*, with myside bias being unaffected by source characteristics. Prior exposure similarly affects truth judgements by reducing overall threshold. Whether prior exposure also affects truth sensitivity and myside bias remains unclear at this point. Finally, more work is needed to determine whether psychological inoculation increases truth sensitivity or overall threshold (or both). The potential impact of inoculation on myside bias remains unclear due to the lack of studies that manipulated the attitude-congruence of the focal information.

Implications

Why do people fall for misinformation?

The reviewed findings pose a challenge to dominant narratives about why people fall for misinformation. First, counter to claims that people are bad at discerning true from false information, the available evidence suggests that truth sensitivity is remarkably high on average (Gawronski et al., 2025). Thus, while truth sensitivity is important for understanding why people mistakenly judge false information as true, the available evidence suggests that low truth sensitivity plays a less significant role than suggested by a dominant narrative in the field.

Second, counter to the idea that people fall for misinformation because of a general tendency to accept information as true (see Brashier & Marsh, 2020; Levine, 2014), the available evidence suggests that, if anything, people show a general tendency to reject information as false. These findings similarly suggest that the presumed general tendency to accept information as true matters much less than suggested by a dominant narrative in the field.

Third, the available evidence suggests that myside bias is pervasive and large on average. Yet, its role in misinformation susceptibility is more complex compared to the unidirectional roles of truth sensitivity and overall threshold. Whereas high truth sensitivity and high overall threshold involve a lower likelihood of judging false information as true for both attitude-congruent and attitude-incongruent information, myside bias involves a stronger susceptibility to attitude-congruent misinformation and weaker susceptibility to attitude-incongruent misinformation. In other words, the misinformation that people accept as true is very likely to be attitude-congruent and very unlikely to be attitude-incongruent.

Two types of errors

The differential outcomes of myside bias in judgements of attitude-congruent and attitude-incongruent information highlight the importance of considering trade-offs between misses (i.e., Type-1 errors) and false alarms (i.e., Type-2 errors) in truth judgements of true and false information. Research on misinformation susceptibility is primarily concerned with false alarms: the erroneous acceptance of false information as true. Yet, if the underlying concern is about incorrect beliefs more broadly, one also needs to consider misses: the erroneous rejection of true information as false. The distinction between the two kinds of errors has been central in debates about the effectiveness of inoculation interventions (see Modirrousta-Galian & Higham, 2023; Simchon et al., 2026), in that an increase in overall threshold potentially caused by these interventions involves not only the intended decrease in judgements of false information as true (i.e., reduced rate of false alarms) but also an unintended increase in judgements of true information as false (i.e., increased rate of misses). The distinction has also been central in discussions of meta-analytic findings suggesting that scepticism towards true information is a greater threat to belief accuracy than gullibility to false information (Pfänder & Altay, 2025).

Our findings paint an even more complex picture, in that gullibility and scepticism differ for attitude-congruent and attitude-incongruent information. To illustrate this point, Figure 8 depicts the score distributions and effect sizes of the two components of myside bias: acceptance threshold for attitude-congruent and attitude-incongruent information. The depicted effect sizes suggest that the tendency to judge attitude-incongruent information as false is three times larger than the tendency to judge attitude-

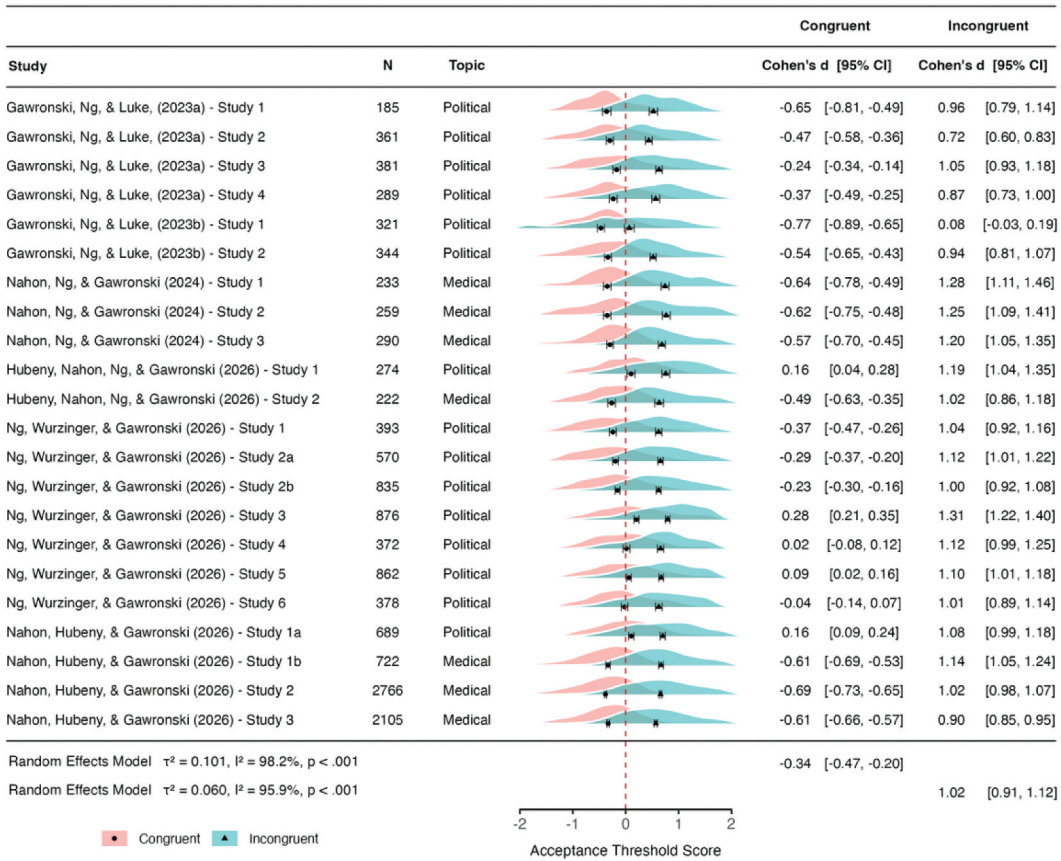

Figure 8. Distributions of acceptance-threshold scores and standardised mean differences (Cohen's *d*) for attitude-congruent information (red) and attitude-incongruent information (blue) across studies. Reference point of zero reflects an equal likelihood of judging information as true or false. Scores higher than zero reflect a tendency to judge information as false. Scores lower than zero reflect a general tendency to judge information as true.

congruent information as true. In other words, participants were much more likely to erroneously reject attitude-incongruent true information than to erroneously accept attitude-congruent false information. Based on these findings, it seems critical (a) to expand the dominant focus on erroneous judgements of false information as true (i.e., false alarms) by including erroneous judgements of true information as false (i.e., misses), and (b) to differentiate between attitude-congruent and attitude-incongruent information in the two kinds of errors.

Addressing extant debates

In addition to highlighting the significance of erroneous judgements of true information as false beyond erroneous judgements of false information as true, misinformation research using SDT provides valuable insights for extant debates about why people fall for misinformation. First, applying SDT to truth judgements of true and false information reconciles the debate between researchers who argue that people fall for misinformation because they fail to engage in analytic thinking (e.g., Pennycook & Rand, 2019) and researchers who argue that people fall for misinformation because they aim to support and protect their partisan identities (e.g., Van Bavel & Pereria, 2018). An analysis in terms of SDT suggests that the two positions are not mutually exclusive, in that the former hypothesis pertains to the determinants of truth sensitivity, whereas the latter hypothesis pertains to the determinants of differential thresholds for attitude-congruent and attitude-incongruent information (Batailler et al., 2022). The reviewed findings indicate that both hypotheses are correct (instead of positive evidence for one implying falsity of the other).

Second, the reviewed research speaks to the ongoing debate about whether partisan bias in judgements of misinformation is driven by identity-protective motivated reasoning or purely cognitive processes following principles of Bayesian inference (for a review, see Ditto et al., 2025). The available evidence suggests that both mechanisms independently contribute to partisan bias in judgements of misinformation.

Third, our SDT framework provides conceptual and empirical clarity for the debate between proponents (e.g., Kahan, 2013) and critics (e.g., Pennycook & Rand, 2021) of the so-called *motivated-system-2 hypothesis*, which states that people utilise their cognitive capabilities to support and protect their partisan views. At the conceptual level, an analysis in terms of SDT clarifies that the motivated-system-2 hypothesis has nothing to do with truth discernment (see Pennycook & Rand, 2019) but instead speaks to the relative difference in thresholds for attitude-congruent versus attitude-incongruent information (Batailler et al., 2022). At the empirical level, research using SDT suggests that, while evidence for the motivated-system

-2 hypothesis is scarce in individual studies, meta-analytic evidence supports its validity (Sultan et al., 2024).[8]

What can be done about it?

The reviewed findings also have important implications for the development and evaluation of interventions to reduce susceptibility to misinformation (Gawronski et al., 2024). The most significant conclusion is that interventions should be tailored to the nature of the focal problem: do people fall for misinformation because they are unable to discern true from false information (i.e., low truth sensitivity), because they have a general tendency to accept information as true (i.e., low overall threshold), or because they have a tendency to accept attitude-congruent information and reject attitude-incongruent information (i.e., myside bias)? Expanding on the outcome of this diagnosis, interventions should be designed to target the psychological underpinnings of the identified factor (see Figure 3). Because the three factors have distinct psychological underpinnings, interventions targeting the underpinnings of one factor may have little impact if misinformation susceptibility is rooted in a different factor.

These conclusions have major implications for the effectiveness of person-centred interventions to reduce misinformation susceptibility. Many extant interventions target psychological processes related to truth sensitivity, such as limited knowledge and other cognitive deficits (for reviews, see Ecker et al., 2022; Kozyreva et al., 2024). Yet, truth sensitivity is remarkably high on average, which suggests that low truth sensitivity plays a less significant role than proclaimed by a dominant narrative in the field. From this perspective, it is not particularly surprising that the impact of these interventions has been found to be very limited (see Fazio et al., 2024; Hoes et al., 2024; Sun et al., 2025). Based on the reviewed findings, interventions targeting the psychological roots of myside bias may be more successful. The reviewed findings further suggest that such interventions would have to target meta-cognitive processes related to belief confidence and motivational processes related to identity-protection. Needless to say, targeting these processes requires approaches that are very different from the ones of interventions targeting cognitive deficits.

Our analysis further suggests that, when evaluating the effectiveness of interventions, it is critical to consider both judgements of false information as true (i.e., false alarms) and judgements of true information as false (i.e., misses). In line with this concern, inoculation interventions have been

[8]Recent work by Lois et al. (2026) suggests that the motivated-system-2 hypothesis may be valid for what can be described as hard news (e.g., information about policy-related issues), but not for soft news (e.g., information about political figures).

criticised for reducing false alarms at the expense of increasing misses (Hoes et al., 2024). This pattern is reflected in increased overall threshold, in that inoculation interventions have been claimed to increase participants' general tendency to judge all information as false (Modirrousta-Galian & Higham, 2023; Sun et al., 2025; but see; Simchon et al., 2026). Given that average levels of overall threshold already involve a general tendency to reject information as false (see Figure 5), further increases in overall threshold may be more detrimental than helpful, because (a) the average rate of misses is higher than the average rate of false alarms, and (b) the interventions further increase the high rate of misses. If this increased scepticism is selectively applied to attitude-incongruent information without being applied to attitude-congruent information (see Loughnan et al., 2026), the outcome could be even more problematic, in that the interventions may inadvertently increase myside bias. While such an outcome would be consistent with the original purpose of inoculation interventions to immunise people against undesired propaganda (McGuire, 1964), it is inconsistent with the purpose of applied misinformation research to foster accurate beliefs (Guay et al., 2023). In general, interventions that aim to increase accurate beliefs should reduce false alarms (i.e., Type-2 errors) without increasing misses (i.e., Type-1 errors), and vice versa.

Similar concerns apply to potential interventions targeting myside bias, which involves a more complex pattern of errors. Overall, myside bias is characterised by (a) high rates of false alarms and low rates of misses in judgements of attitude-congruent information and (b) high rates of misses and low rates of false alarms in judgements of attitude-incongruent information. The reviewed findings further suggest that scepticism against attitude-incongruent true information is much more pronounced than gullibility to attitude-congruent false information (see Figure 8), indicating that people are more likely to erroneously judge attitude-incongruent true information as false than to erroneously judge attitude-congruent false information as true (see Table 2). Thus, if one is concerned about incorrect beliefs more broadly, it seems critical to tackle not only gullibility to attitude-congruent false information, but also scepticism towards attitude-incongruent true information. Moreover, to qualify as effective, interventions should ideally reduce false alarms without increasing misses in judgements of attitude-congruent information, and they should reduce misses without increasing false alarms in judgements of attitude-incongruent information (see Table 2). Arguably, such interventions will require approaches that are fundamentally different from the ones of extant interventions that aim to tackle the cognitive deficits presumed to underlie misinformation susceptibility.

More broadly, our analysis suggests a four-step approach for the development and evaluation of person-centred misinformation interventions (see Table 3; Gawronski et al., 2024). The first step involves diagnosing the nature

Table 2. Proportions of false information judged as true (false alarms) and true information judged as false (misses) in truth judgements of attitude-congruent and attitude-incongruent information.

	Topic	*N*	False Alarms		Misses	
			Congruent	Incongruent	Congruent	Incongruent
Gawronski et al. (2023a) Study 1	Political	185	.52	.23	.28	.57
Gawronski et al. (2023a) Study 2	Political	361	.51	.29	.32	.56
Gawronski et al. (2023a) Study 3	Political	381	.43	.20	.33	.59
Gawronski et al. (2023a) Study 4	Political	289	.48	.24	.33	.60
Gawronski et al. (2023b) Study 1	Political	321	.58	.42	.29	.47
Gawronski et al. (2023b) Study 2	Political	344	.53	.25	.31	.59
Nahon et al. (2024) Study 1	Medical	233	.36	.09	.16	.51
Nahon et al. (2024) Study 2	Medical	259	.37	.10	.17	.52
Nahon et al. (2024) Study 3	Medical	290	.34	.10	.17	.48
Hubeny, Nahon, Ng, et al. (2026) Study 1	Political	274	.38	.18	.44	.63
Hubeny, Nahon, Ng, et al. (2026) Study 2	Medical	222	.35	.13	.19	.48
Ng et al. (2026) Study 1	Political	393	.43	.18	.28	.56
Ng et al. (2026) Study 2a	Political	570	.43	.19	.31	.59
Ng et al. (2026) Study 2b	Political	835	.43	.20	.33	.59
Ng et al. (2026) Study 3	Political	876	.34	.15	.47	.66
Ng et al. (2026) Study 4	Political	372	.39	.18	.40	.60
Ng et al. (2026) Study 5	Political	862	.37	.18	.41	.61
Ng et al. (2026) Study 6	Political	378	.41	.21	.40	.60
Nahon et al. (2026) Study 1a	Political	689	.37	.18	.43	.60
Nahon et al. (2026) Study 1b	Medical	722	.35	.10	.16	.47
Nahon et al. (2026) Study 2	Medical	2,766	.39	.12	.16	.48
Nahon et al. (2026) Study 3	Medical	2,105	.36	.13	.17	.45
Sample-weighted Average		13,727	.40	.16	.27	.54

Table 3. Recommended steps for the development and evaluation of person-centred misinformation interventions.

Step 1	Diagnosis of Problem	Diagnose whether people fall for misinformation due to low truth sensitivity, low overall threshold, or myside bias.
Step 2	Identification of Psychological Roots	Identify psychological underpinnings of problem identified at first step.
Step 3	Development of Intervention	Develop intervention that targets the psychological roots identified at second step.
Step 4	Evaluation of Intervention	Evaluate if intervention reduces prevalent types of errors in problem identified at first step without increasing other types of errors.

of the problem: do people fall for misinformation because they are unable to discern true from false information (i.e., low truth sensitivity), because they have a general tendency to accept information as true (i.e., low overall threshold), or because they have a tendency to accept attitude-congruent information and reject attitude-incongruent information (i.e., myside bias)? The second step involves identifying the psychological roots of the focal problem identified at the first step: why do people show low truth sensitivity, low overall threshold, or myside bias? The third step involves the development of interventions that target the psychological roots identified at the second step. Finally, the fourth step involves evaluating the interventions developed at the third step in terms of whether they effectively reduce the most prevalent types of errors in the problem identified at the first step without increasing other types of errors.

Positionality and constraints on generality

The first author and his lab members became interested in psychological research on misinformation when they discussed a preprint of Pennycook and Rand (2019) during a lab meeting. Everyone in the lab was surprised by the authors' claim that "susceptibility to fake news is driven more by lazy thinking than it is by partisan bias" (p. 39). Yet, when discussing Pennycook and Rand's results, several lab members expressed concerns that (a) the proclaimed lack of partisan bias may be due to Pennycook and Rand's use of a data-analytic approach that is not well suited to identify partisan bias and (b) SDT would be superior to gauge the role of partisan bias in judgements of real and fake news. In line with these concerns, a reanalysis of Pennycook and Rand's data using SDT revealed that partisan bias is by far the largest effect in their data, in that participants showed a significantly lower acceptance threshold for attitude-congruent than attitude-incongruent information (Batailler et al., 2022). The results of this reanalysis inspired the first author and three of his lab members to write a methodological article on the value of SDT for studying the identification of fake news (Batailler et al., 2022), which included the results of their reanalysis of Pennycook and Rand's

data. This paper became the basis for the research programme reviewed in the current article, which confirmed our preconception that partisan bias (or myside bias in a broader sense) is critical for understanding susceptibility to misinformation.

While we deem it important to disclose our preconception about the significance of myside bias in judgements of true and false information, we should also mention other preconceptions that were not confirmed by our data. First, when we started our research on misinformation susceptibility, we were convinced that people are not very good at discerning true from false information (see Gawronski et al., 2025). This assumption turned out to be false, in that participants showed remarkably high levels of truth sensitivity (see Figure 4). Second, we were convinced that people have a low overall threshold for judging information as true (see Gawronski et al., 2025). This assumption also turned out to be false, in that participants showed a general tendency to judge information as false instead of a general tendency to judge information as true (see Figure 5).

Regarding generalisability, we are convinced that SDT provides a valuable framework for understanding truth judgements of true and false information, irrespective of the empirically obtained contributions of truth sensitivity, overall threshold, and myside bias. The framework itself is agnostic about which factors are more or less important; it simply provides a tool to study their respective contributions. In this sense, we consider the framework as generalisable, even when the findings obtained with the framework may differ across people, contexts, time, or cultures. Yet, the latter possibility is critical for the generalisability of our conclusions about average levels of truth sensitivity, overall threshold, and myside bias at the first level of analysis. Because the reviewed studies almost exclusively relied on participants from the United States who were recruited via Prolific,[9] the generalisability of our conclusions regarding the first level of analysis remains to be tested. It is certainly possible that, different from the reported results, other populations show lower levels of truth sensitivity, a lower overall threshold involving a general tendency to judge information as true, or lower levels of myside bias. Future research is needed to address these questions.

As for our arguments pertaining to the second level of analysis, we consider our conclusions generalisable, in that the psychological underpinnings of truth sensitivity, overall threshold, and myside bias can be assumed to be identical across people, contexts, time, and cultures. While there may be population-related differences in the overall levels of a given construct

[9]A notable exception are four studies on misinformation about Covid-19 vaccines that included participants from the United Kingdom (Hubeny, Nahon, Ng, et al., 2026; Nahon et al., 2024).

(e.g., average levels of myside bias may differ across countries), we are convinced that the obtained relations between constructs (e.g., higher levels of confidence being associated with greater myside bias) reflect universal principles of psychological functioning (e.g., overconfidence increasing myside bias). Yet, any such claims require empirical data to gauge whether the obtained relations replicate in different populations and, if not, follow-up investigations to determine whether the proclaimed universal principles can explain the obtained differences across populations.

A more substantial caveat about generalisability pertains to the fact that – different from the focus on objective stimulus properties in the original application of SDT to psychophysics – truth is not an objective property of stimuli but essentially depends on the background knowledge of the perceiver (Quine, 1953). This aspect has important implications for the generalisability of our conclusions, because it means that any finding obtained with our SDT framework is jointly shaped by characteristics of the participants and the employed stimuli. This caveat applies to research at both levels of analysis. For example, while truth sensitivity was remarkably high in the reviewed studies, it seems unlikely that participants from countries or regions with little exposure to U.S. American politics would show comparable levels of truth sensitivity in judging true and false information about American politics (first level of analysis). In the absence of domain-relevant knowledge, greater cognitive elaboration may also be ineffective in increasing truth sensitivity (second level of analysis). Moreover, even for participants from the United States with a certain level of knowledge about U.S. American politics, truth sensitivity will likely vary across stimulus sets with different levels of item difficulty. Another methodological issue is that all misinformation studies using SDT require a decision about the relative proportion of true versus false information. Thus, if the (usually equal) proportions in the reviewed studies differ from the (presumably unequal) proportions in natural settings, participants may adjust their overall threshold if they notice the difference. In this case, the obtained general tendency to reject information as false may be limited to settings with balanced proportions of true and false information; it may not generalise to settings with higher proportions of true information, as might be expected of natural information ecologies.

Conclusion

A framework based on SDT suggests that truth judgements of true and false information are shaped by three factors: (a) accurate discernment of true and false information (*truth sensitivity*); (b) a general tendency to judge information as true versus false (*overall threshold*); and (c) a tendency to accept attitude-congruent information and to reject attitude-incongruent information (*myside bias*). The reviewed findings

indicate that, while all three factors matter, myside bias is the most significant factor for understanding why people fall for misinformation and what can be done about it. Thus, interventions to reduce susceptibility to misinformation will likely be most effective if they target the meta-cognitive and motivational processes underlying myside bias instead of processes underlying truth sensitivity or overall threshold. Moreover, if the concern is about inaccurate beliefs more broadly, interventions should adopt a broader focus that includes not only erroneous judgements of attitude-congruent false information as true but also erroneous judgements of attitude-incongruent true information as false.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Science Foundation [BCS-2040684], the Swiss National Science Foundation [P500PS_214298 and P5R5PS_225565], and the John Templeton Foundation [62824]. 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 funding agencies. The funders had no role in the decision to publish or the preparation of this manuscript.

ORCID

Bertram Gawronski http://orcid.org/0000-0001-7938-3339
Nyx L. Ng http://orcid.org/0000-0002-3416-1385
Lea S. Nahon http://orcid.org/0000-0003-2468-4243
Tyler J. Hubeny http://orcid.org/0009-0008-7069-4593

Data availability statement

The analysis codes and information on the availability of the data are available at https://osf.io/n5tbv/.

References

Ames, D. R., Rose, P., & Anderson, C. P. (2006). The NPI-16 as a short measure of narcissism. *Journal of Research in Personality*, *40*(4), 440–450. https://doi.org/10.1016/j.jrp.2005.03.002

Barajas, J., & John, R. (2023). A signal detection theory approach to predicting immunity to pandemic vaccine fake news. *Proceedings of the 56th Hawaii International Conference on System Sciences*, Article 9. https://aisel.aisnet.org/hicss-56/dg/disaster_resilience/9

Basol, M., Roozenbeek, J., & van der Linden, S. (2020). Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition*, *3*(1), Article 2. https://doi.org/10.5334/joc.91

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

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

Bruder, M., Haffke, P., Neave, N., Nouripanah, N., & Imhoff, R. (2013). Measuring individual differences in generic beliefs in conspiracy theories across cultures: Conspiracy mentality questionnaire. *Frontiers in Psychology*, *4*, Article 43078. https://doi.org/10.3389/fpsyg.2013.00225

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., Celniker, J. B., Siddiqi, S. S., Güngör, M., & Relihan, D. P. (2025). Partisan bias in political judgment. *Annual Review of Psychology*, *76*(1), 717–740. https://doi.org/10.1146/annurev-psych-030424-122723

Druckman, J. N., & McGrath, M. C. (2019). The evidence for motivated reasoning in climate change preference formation. *Nature Climate Change*, *9*(2), 111–119. https://doi.org/10.1038/s41558-018-0360-1

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Amazeen, M. A., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13–29. https://doi.org/10.1038/s44159-021-00006-y

Ecker, U. K. H., Tay, L. Q., Roozenbeek, J., van der Linden, S., Cook, J., Oreskes, N., & Lewandowsky, S. (2025). Why misinformation must not be ignored. *American Psychologist*, *80*(6), 867–878. https://doi.org/10.1037/amp0001448

Fazio, L., Rand, D., Lewandowsky, S., Susmann, M., Berinsky, A. J., Guess, A., & Swire-Thompson, B. (2024). *Combating misinformation: A megastudy of nine interventions designed to reduce the sharing of and belief in false and misleading headlines* [Unpublished manuscript]. https://doi.org/10.31234/osf.io/uyjha

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

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., & Bodenhausen, G. V. (2015). Theory evaluation. In B. Gawronski & G. V. Bodenhausen (Eds.), *Theory and explanation in social psychology* (pp. 3–23). Guilford Press.

Gawronski, B., Nahon, L. S., & Ng, N. L. (2024). A signal-detection framework for misinformation interventions. *Nature Human Behaviour*, *8*(12), 2272–2274. https://doi.org/10.1038/s41562-024-02021-4

Gawronski, B., Nahon, L. S., & Ng, N. L. (2025). Debunking three myths about misinformation. *Current Directions in Psychological Science*, *34*(1), 36–42. https://doi.org/10.1177/09637214241280907

Gawronski, B., Ng, N. L., & Luke, D. (2023b). *Two failed attempts to influence partisan bias in responses to misinformation via self-affirmation* [Unpublished manuscript]. https://doi.org/10.31219/osf.io/qr8xb

Gawronski, B., Ng, N. L., & Luke, D. M. (2023a). Truth sensitivity and partisan bias in responses to misinformation. *Journal of Experimental Psychology General*, *152* (8), 2205–2236. https://doi.org/10.1037/xge0001381

Grant, M. D., Markowitz, D. M., Sherman, D. K., Flores, A., Dickert, S., Eom, K., Van Boven, L., Kogut, T., Mayorga, M., Oonk, D., Pedersen, E. J., Pereira, B., Rubaltelli, E., Slovic, P., Västfjäll, D., & Van Boven, L. (2024). Ideological diversity of media consumption predicts COVID-19 vaccination. *Scientific Reports*, *14*(1), Article 28948. https://doi.org/10.1038/s41598-024-77408-4

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

Guay, B., Berinsky, A. J., Pennycook, G., & Rand, D. (2023). How to think about whether misinformation interventions work. *Nature Human Behaviour*, *7*(8), 1231–1233. https://doi.org/10.1038/s41562-023-01667-w

Hoes, E., Aitken, B., Zhang, J., Gackowski, T., & Wojcieszak, M. (2024). Prominent misinformation interventions reduce misperceptions but increase scepticism. *Nature Human Behaviour*, *8*(8), 1545–1553. https://doi.org/10.1038/s41562-024-01884-x

Hubeny, T. J., Nahon, L. S., & Gawronski, B. (2026). Understanding partisan bias in judgments of misinformation: Identity protection versus differential knowledge. *Psychological Science*, *37*(1), 43–54. https://doi.org/10.1177/09567976251404040

Hubeny, T. J., Nahon, L. S., Ng, N. L., & Gawronski, B. (2026). Who falls for misinformation and why? *Personality & Social Psychology Bulletin*. https://doi.org/10.1177/01461672251328800

Iyengar, A., Gupta, P., & Priya, N. (2023). Inoculation against conspiracy theories: A consumer side approach to India's fake news problem. *Applied Cognitive Psychology*, *37*(2), 290–303. https://doi.org/10.1002/acp.3995

Jarvis, W. B. G., & Petty, R. E. (1996). The need to evaluate. *Journal of Personality & Social Psychology*, *70*(1), 172–194. https://doi.org/10.1037/0022-3514.70.1.172

Jiang, Y., Schwarz, N., Reynolds, K. J., & Newman, E. J. (2024). Repetition increases belief in climate-skeptical claims, even for climate science endorsers. *PLOS ONE*, *19*(8), Article e0307294. https://doi.org/10.1371/journal.pone.0307294

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment & Decision Making*, *8*(4), 407–424. https://doi.org/10.1017/S1930297500005271

Kelley, H. H., & Michela, J. L. (1980). Attribution theory and research. *Annual Review of Psychology*, *31*(1), 457–501. https://doi.org/10.1146/annurev.ps.31.020180.002325

Koetke, J., Schumann, K., Porter, T., & Smilo-Morgan, I. (2023). Fallibility salience increases intellectual humility: Implications for people's willingness to investigate political misinformation. *Personality & Social Psychology Bulletin*, *49*(5), 806–820. https://doi.org/10.1177/01461672221080979

Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K. H., Lewandowsky, S., Hertwig, R., Wineburg, S., Bak-Coleman, J., Barzilai, S., Basol, M., Berinsky, A. J., Betsch, C., Cook, J., Fazio, L. K., Geers, M., Guess, A. M., Huang, H., Larreguy, H., Maertens, R. . . . van der Linden, S. (2024). Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, *8*(6), 1044–1052. https://doi.org/10.1038/s41562-024-01881-0

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., Zittrain, J. L., 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. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Leary, M. R., Diebels, K. J., Davisson, E. K., Jongman-Sereno, K. P., Isherwood, J. C., Raimi, K. T., Hoyle, R. H., & Hoyle, R. H. (2017). Cognitive and interpersonal features of intellectual humility. *Personality & Social Psychology Bulletin*, *43*(6), 793–813. https://doi.org/10.1177/0146167217697695

Leary, M. R., Kelly, K. M., Cottrell, C. A., & Schreindorfer, L. S. (2013). Construct validity of the need to belong scale: Mapping the nomological network. *Journal of Personality Assessment*, *95*(6), 610–624. https://doi.org/10.1080/00223891.2013.819511

Levine, T. R. (2014). Truth-default theory (TDT): A theory of human deception and deception detection. *Journal of Language & Social Psychology*, *33*(4), 378–392. https://doi.org/10.1177/0261927X14535916

Lewandowsky, S., Ecker, U. K., 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

Lois, G., Gardikiotis, A., Karaspyrou, Z., Tsakanikos, E., & Sedikides, C. (2026). Rethinking the link between cognitive reflection and susceptibility to political misinformation: Distinguishing hard from soft news. *Political Psychology*, *47*(2), Article e70109. https://doi.org/10.1111/pops.70109

Loughnan, D., van Stekelenburg, A., Pouwels, J. L., Fransen, M. L., & Kleemans, M. (2026). An analysis of studies testing digital interventions to inoculate against misinformation: A systematic review. *Communication Research*. https://doi.org/10.1177/00936502251411467

Ludwig, J., & Sommer, J. (2024). Mindsets and politically motivated reasoning about fake news. *Motivation and Emotion*, *48*(3), 249–263. https://doi.org/10.1007/s11031-024-10067-0

Lyons, B., Modirrousta-Galian, A., Altay, S., & Salovich, N. (2025). Reducing blind spots? Performance feedback reduces relative confidence but does not improve subsequent news discernment. *Collabra: Psychology*, *11*(1), Article 138652. https://doi.org/10.1525/collabra.138652

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, *27*(1), 1–16. https://doi.org/10.1037/xap0000315

McFarland, S., Webb, M., & Brown, D. (2012). All humanity is my ingroup: A measure and studies of identification with all humanity. *Journal of Personality & Social Psychology*, *103*(5), 830–853. https://doi.org/10.1037/a0028724

McGuire, W. J. (1964). Inducing resistance to persuasion: Some contemporary approaches. *Advances in Experimental Social Psychology*, *1*, 191–229. https://doi.org/10.1016/S0065-2601(08)60052-0

Modirrousta-Galian, A., & Higham, P. A. (2023). Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with receiver operating characteristic analysis. *Journal of Experimental Psychology General*, *152*(9), 2411–2437. https://doi.org/10.1037/xge0001395

Mussweiler, T. (2003). Comparison processes in social judgement: Mechanisms and consequences. *Psychological Review*, *110*(3), 472–489. https://doi.org/10.1037/0033-295X.110.3.472

Nahon, L. S., Hubeny, T. J., & Gawronski, B. (2026). *Does fallibility salience reduce myside bias in judgments of misinformation?* Unpublished manuscript.

Nahon, L. S., Ng, N. L., & Gawronski, B. (2024). Susceptibility to misinformation about COVID-19 vaccines: A signal detection analysis. *Journal of Experimental Social Psychology*, *114*, Article 104632. https://doi.org/10.1016/j.jesp.2024.104632

Ng, N. L., Wurzinger, F. M., & Gawronski, B. (2026). *A signal detection analysis of source effects on susceptibility to misinformation* Unpublished manuscript.

Pennycook, G. (2023). A framework for understanding reasoning errors: From fake news to climate change and beyond. *Advances in Experimental Social Psychology*, *67*, 131–208. https://doi.org/10.1016/bs.aesp.2022.11.003

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases the perceived accuracy of fake news. *Journal of Experimental Psychology General*, *147* (12), 1865–1880. https://doi.org/10.1037/xge0000465

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015). On the reception and detection of pseudo-profound bullshit. *Judgment & Decision Making*, *10*(6), 549–563. https://doi.org/10.1017/S1930297500006999

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. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, *88*(2), 185–200. https://doi.org/10.1111/jopy.12476

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

Perez Santangelo, A., & Solovey, G. (2023). Understanding belief in political statements using a model-driven experimental approach: A registered report. *Scientific Reports*, *13*(1), Article 21205. https://doi.org/10.1038/s41598-023-47939-3

Pfänder, J., & Altay, S. (2025). Spotting false news and doubting true news: A systematic review and meta-analysis of news judgements. *Nature Human Behaviour*, *9*(4), 688–699. https://doi.org/10.1038/s41562-024-02086-1

Porter, T. (2025). Intellectual humility: On recognizing our limits. *Advances in Experimental Social Psychology*, *71*, 135–179. https://doi.org/10.1016/bs.aesp.2024.10.005

Porter, T., Elnakouri, A., Meyers, E. A., Shibayama, T., Jayawickreme, E., & Grossmann, I. (2022). Predictors and consequences of intellectual humility. *Nature Reviews Psychology*, *1*(9), 524–536. https://doi.org/10.1038/s44159-022-00081-9

Quine, W. V. O. (1953). Two dogmas of empiricism. In W. V. O. Quine (Ed.), *From a logical point of view* (pp. 20–46). Harvard University Press.

Ramos, G. A., & Van Boven, L. (2025). The age of misinformation: Older people exhibit greater partisan bias in sharing and evaluating (mis)information accuracy. *Journal of Experimental Psychology General*, *155*(1), 197–214. https://doi.org/10.1037/xge0001868

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–150. https://doi.org/10.1037/met0000045

Roozenbeek, J., Maertens, R., McClanahan, W., & van der Linden, S. (2020). Disentangling item and testing effects in inoculation research on online misinformation: Solomon revisited. *Educational and Psychological Measurement*, *81*(2), 340–362. https://doi.org/10.1177/0013164420940378

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), Article 65. https://doi.org/10.1057/s41599-019-0279-9

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton University Press.

Schwarz, N., Sanna, L., 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

Seabrooke, T., Modirrousta-Galian, A., & Higham, P. A. (2026). Re-examining the bad news game: No evidence of improved discrimination of Indian true and fake news headlines. *Psychonomic Bulletin and Review*, *33*(1), Article 13. https://doi.org/10.3758/s13423-025-02827-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

Simchon, A., Zipori, T., Teitelbaum, L., Lewandowsky, S., & van der Linden, S. A. (2026). Signal detection theory meta-analysis of psychological inoculation against misinformation. *Current Opinion in Psychology*, *67*, Article 102194. https://doi.org/10.1016/j.copsyc.2025.102194

Soto, C. J., & John, O. P. (2017). Short and extra-short forms of the big five inventory-2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality*, *68*, 69–81. https://doi.org/10.1016/j.jrp.2017.02.004

Stanovich, K. E., & Toplak, M. E. (2023). Actively open-minded thinking and its measurement. *Journal of Intelligence*, *11*(2), Article 2. https://doi.org/10.3390/jintelligence11020027

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, *22*(4), 259–264. https://doi.org/10.1177/0963721413480174

Sultan, M., Tump, A. N., Ehmann, N., Lorenz-Spreen, P., Hertwig, R., Gollwitzer, A., & Kurvers, R. H. J. M. (2024). Susceptibility to online misinformation: A systematic meta-analysis of demographic and psychological factors. *Proceedings of the National Academy of Sciences*, *121*(47), Article e2409329121. https://doi.org/10.1073/pnas.2409329121

Sultan, M., Tump, A. N., Geers, M., Lorenz-Spreen, P., Herzog, S. M., & Kurvers, R. H. (2022). Time pressure reduces misinformation discrimination ability but does not alter response bias. *Scientific Reports*, *12*(1), Article 22416. https://doi.org/10.1038/s41598-022-26209-8

Sun, X., Bai, X., Chen, B., Niu, G., & Mao, P. (2025). *The impact of prebunking interventions against misinformation on discrimination ability and criterion: An IPD network meta-analysis* [Unpublished manuscript]. https://doi.org/10.21203/rs.3.rs-6660774/v1

Tappin, B. M., Pennycook, G., & Rand, D. G. (2020). Thinking clearly about causal inferences of politically motivated reasoning: Why paradigmatic study designs often undermine causal inference. *Current Opinion in Behavioral Sciences*, *34*, 81–87. https://doi.org/10.1016/j.cobeha.2020.01.003

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment & Decision Making*, *11*(1), 99–113. https://doi.org/10.1017/S1930297500007622

Udry, J., & Barber, S. J. (2024). The illusory truth effect: A review of how repetition increases belief in misinformation. *Current Opinion in Psychology*, *56*, Article 101736. https://doi.org/10.1016/j.copsyc.2023.101736

Unkelbach, C., Koch, A., Silva, R. R., & Garcia-Marques, T. (2019). Truth by repetition: Explanations and implications. *Current Directions in Psychological Science*, *28*(3), 247–253. https://doi.org/10.1177/0963721419827854

Van Bavel, J. J., & Pereria, 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

Van Bavel, J. J., Rathje, S., Vlasceanu, M., & Pretus, C. (2024). Updating the identity-based model of belief: From false belief to the spread of misinformation. *Current Opinion in Psychology*, *56*, Article 101787. https://doi.org/10.1016/j.copsyc.2023.101787

Van der Linden, S. (2024). Countering misinformation through psychological inoculation. *Advances in Experimental Social Psychology*, *69*, 1–58. https://doi.org/10.1016/bs.aesp.2023.11.001

Van der Linden, S., Albarracín, D., Fazio, L., Freelon, D., Roozenbeek, J., Swire-Thompson, B., & Van Bavel, J. (2025). Using psychological science to understand and fight health misinformation: An APA consensus statement. *American Psychologist*. https://doi.org/10.1037/amp0001598

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03