Facts over partisanship: Evidence-based updating of trust in partisan sources

Giannis Lois[1,2], Elias Tsakas[2], Arno Riedl[2]

[1]Department of Psychology, School of Social Sciences, University of Crete, 74150 Rethymno, Greece

[2]Department of Microeconomics and Public Economics, School of Business and Economics, Maastricht University, 6200 MD Maastricht, The Netherlands

Corresponding author: Giannis Lois

Email address: i.lois@uoc.gr

Word count: 9413

Abstract

A prominent explanation for the proliferation of political misinformation and the growing belief polarization is that people engage in motivated reasoning to affirm their ideology and to protect their political identities. An alternative explanation is that people seek the truth but use partisanship as a heuristic to discern credible from dubious sources of political information. In two experiments, we test these competing explanations in a dynamic setting where Democrats and Republicans are repeatedly exposed to messages from ingroup or outgroup partisan sources and can gradually learn which source is credible based on external feedback. Both Democrats and Republicans initially incorporated information from ingroup sources more than from outgroup sources. This pattern was stronger among partisans that displayed high affective polarization. Across rounds, this partisan bias declined, or even changed direction, as supporters of both groups gradually incorporated information from reliable sources more than unreliable sources irrespective of the source's partisanship. Importantly, the content of the shared information (i.e., neutral vs political) and the presence of partisan sources as opposed to neutral sources did not affect the learning process indicating the presence of strong accuracy motives. In contrast, increased uncertainty regarding source reliability undermined the learning process. These findings demonstrate that partisans follow Bayesian learning dynamics. Although they initially display a partisan bias in the incorporation of information, they overcome this bias in the presence of external feedback and learn to trust credible sources irrespective of partisanship.

Keywords: belief updating, source reliability, trust, partisan bias, directional reasoning

Introduction

In the last decade, there has been widespread concern about the proliferation of misinformation and the growing trend of belief polarization which have extended to factual issues such as human-caused climate change (Rutjens et al., 2018), vaccines safety (Fridman et al., 2021), or the outcome of the 2020 US presidential election (Kahn 2021). More troubling is the fact that misinformed and polarized views persist even after repeated efforts to debunk them through fact-based messages (Flynn et al., 2017; Taber & Lodge, 2006).

Cognitive psychology offers two plausible explanations for this resistance to evidence and the growing belief polarization. According to a "directional reasoning" account, people strive to protect their ideology (Feinberg & Willer, 2013; Wolsko et al., 2016) or their valuable social identities (Kahan, 2016; Turner et al., 1994) and thus evaluate new information in ways that affirm their prior or desired beliefs (Cohen, 2003; Kahan, 2013; Xiao et al., 2016). This process leads to deviations from rational (Bayesian) updating and is particularly prevalent in political contexts where identity-protective motives are very salient (Druckman & McGrath, 2019; Kahan, 2016). Nonetheless, directional reasoning can also apply to neutral, identity-irrelevant beliefs when group polarization is extreme and information comes from partisan sources (Abrams et al., 2003). For instance, a recent survey showed that the majority of Democrats and Republicans not only disagree over policy-related issues but also cannot agree on basic facts such as the size of a demonstration (Laloggia, n.d.).

An alternative "accuracy motives with biased priors" account posits that biased incorporation of new information and the resulting belief polarization can be consistent with Bayesian reasoning. According to this account, erroneous or polarized beliefs are the product of a procedurally rational process in which accuracy-motivated individuals selectively incorporate information from sources that they deem credible based on prior impressions and prejudice rather than actual source

credibility(Clemm Von Hohenberg & Guess, 2023; Druckman & McGrath, 2019). This mechanism is particularly relevant in political contexts in which partisanship serves as a heuristic to discern credible from dubious sources of information. Given that partisan sources predominantly share identity-congruent information (Hansen & Kim, 2011; Pronin et al., 2002; Swire et al., 2017; Van Bavel & Pereira, 2018), a tendency to endorse information from co-partisan sources and to discount information from rival partisan sources will inevitably lead to polarized beliefs and attitudes.

Existing work has failed to disentangle the "directional reasoning" from the "accuracy motives with biased priors" mechanism as they both lead to biased beliefs in the short-term (Druckman & McGrath, 2019). However, under certain conditions, the two mechanisms result in substantially different long-term outcomes. A good illustration of this difference is Republican supporters' trust towards Donald Trump before and after his presidency. During his presidency, Donald Trump made numerous misleading or false claims that were repeatedly debunked by mainstream media (Baker, 2018; Swire et al., 2017; The Washington Post, 2021). Accuracy-motivated Republicans may have initially trusted Trump based on pre-existing stereotypes or heuristics about the credibility and trustworthiness of Republicans politicians. However, in light of the extensive debunking of his misleading and false claims, accuracy-motivated partisans would gradually reduce their trust in Trump's statements. On the other hand, partisans who consistently engage in directional reasoning would maintain and even strengthen their trust in Trump given that he repeatedly shared identity-congruent (mis)information.

Current Study

The present study constitutes the first experimental attempt to disentangle these two mechanisms by using a dynamic setting where participants can gradually update their initial impressions about the credibility of partisan information sources based on evidence. Unlike previous studies that

used self-reported measures of trust (Hansen & Kim, 2011; Vallone et al., 1985; Van Bavel & Pereira, 2018), we used a behavioral incentivized measure that operationalizes trust as the extent to which partisans incorporate information from different sources into their prior beliefs (Schulz et al., 2023). This measure allows us to avoid social desirability biases or partisan cheerleading (Bullock et al., 2013; Peterson & Iyengar, 2021).

We developed a novel belief updating task in which partisans can update their prior estimation about a visual stimulus after receiving a message from a supporter of their political party (i.e., ingroup) or the opposing political party (i.e., outgroup). Subsequently, they receive feedback about the correct answer. After repeated exposure to messages from the same sources and given the presence of external feedback, accuracy-motivated individuals, but not those engaging in directional reasoning, will gradually learn to trust credible partisan sources even when the information contradicts their prior beliefs and comes from an outgroup member.

Previous research suggests that people often possess both accuracy and identity-protective motives and express them to a different extent depending on the context (Van Bavel & Pereira, 2018). For instance, partisan identities are often activated in political contexts (Grace et al., 2008; Tajfel & Turner, 2004). However, other evidence suggests that ingroup bias is present even in neutral, identity-irrelevant topics (Abrams et al., 2003). To explore the conditions that favor one motive over the other, we examined whether sharing information about a polarizing political topic, compared to neutral information, slows down the learning process. Notably, political content may also introduce a desirability bias as partisans tend to endorse ideology-congruent information and discount ideology-incongruent information irrespective of the identity of the source. To further explore whether accuracy and identity-protective motives are simultaneously present, we introduced a control condition in which messages come from neutral (non-partisan) sources. This

condition allows us to measure the baseline rate of learning in the absence of directional reasoning.

Noise is inherent in real-world environments as information sources are very rarely characterized by absolute (in)accuracy and external feedback is accurate with some given probability. Previous evidence suggests that noisy environments aggravate the process of learning and facilitate motivated reasoning and the expression of various biases (Dana et al., 2007; Di Tella et al., 2015; Hamman et al., 2010). To better understand the role of noise in favoring one over the other mechanism, we manipulated the presence and strength of external feedback (Experiment 1) and the degree of uncertainty about actual source credibility (Experiment 2).

Existing work

This work builds on insights from a broader literature on the formation and updating of impressions about others (Hackel et al., 2020; M. Kim et al., 2020; Schulz et al., 2023). According to this literature, people make inferences about others based on heuristics and stereotypes (Dovidio et al., 2010; Fiske et al., 2002; Stanley et al., 2011), or prior selective exposure to information (Derreumaux et al., 2022; Levendusky, 2013; Mothes & Ohme, 2019). These trait inferences can dynamically change based on feedback through a reinforcement learning process (Hackel et al., 2020; M. Kim et al., 2020). Interestingly, this learning process takes place even when people hold strong prior impressions that contradict evidence (M. J. Kim et al., 2021; Leong & Zaki, 2018; Park et al., 2021; Traast et al., 2023). Previous work has documented this evidence-based updating of impressions about others' moral character (Hackel et al., 2020; M. J. Kim et al., 2021; Mende-Siedlecki, 2018) or trustworthiness during economic transactions (Traast et al., 2023).

Our study extends this work by investigating whether partisans update their trust towards ingroup and outgroup information partisan sources based on evidence. In this respect, our work also draws from a surging Behavioral Economics and Political Science literature on the updating of

politically motivated beliefs (Thaler, 2020; Zimmermann, 2020), and the biased seeking of ideologically-aligned sources (Charness et al., 2021). Particularly relevant to our study is a recently proposed theoretical model which showed that small differences in attitude across agents lead to large biases in trust towards these sources after repeated exposure to information from ideologically opposing sources (Gentzkow et al., n.d.).

Hypotheses

We define an "early partisan bias" in trust as the tendency to incorporate information from ingroup sources more than outgroup sources in the first part of the experiment (i.e., before extensive exposure to evidence). First, we formulate hypotheses about this early partisan bias (**H1**):

H1a. Consistent with both the "directional reasoning" account and the "accuracy motives with biased priors" account, participants will exhibit an early partisan bias in trust towards ingroup and outgroup sources. This early partisan bias may reflect identity-protective motives or biased credibility impressions about partisan sources.

H1b. This early partisan bias will be stronger when the content of information is political rather than neutral as identity-relevant content activates partisan identities which strengthens the biased credibility impressions and promotes directional reasoning.

H1c. An alternative hypothesis is that the early partisan bias will be stronger in the face of neutral rather than political content. In the face of political information, partisans' attention will shift from the identity of the sources to the ideological congruence of the information. In other words, partisans will exhibit a desirability bias based on information content irrespective of source identity.

Second, to disentangle the two aforementioned accounts, we formulate hypotheses about the persistence of partisan bias across time and the process of gradually learning to trust reliable sources more than unreliable sources (**H2**):

H2a. Based on the "directional reasoning" account, the partisan bias will persist over time as partisans will not update their trust towards the different sources based on evidence.

H2b. Alternatively, if partisans are motivated by accuracy but have biased prior credibility impressions, the early partisan bias will decline over time (and may even change direction) as partisans will gradually trust reliable sources more than unreliable sources irrespective of partisanship.

H2c. The gradual decline of the partisan bias and the evidence-based updating of trust will be less pronounced in noisy environments in which external evidence is weak (Experiment 1) or there is a high degree of uncertainty about the actual reliability of the sources (Experiment 2).

H2d. The gradual decline of the partisan bias and the evidence-based updating of trust will be less pronounced when the content of information is political rather than neutral. This effect may reflect the more salient identity-protective motives in a political context but may also reflect the presence of a desirability bias which slows down the learning process.

Third, we formulate hypotheses to examine whether the learning process differs depending on whether the identity of the source is partisan or neutral (**H3**):

H3a. Based on the "accuracy motives with biased priors" account, partisans will learn to trust the reliable sources more than the unreliable sources equally fast when sources have partisan and neutral identities.

H3b. In contrast, based on the "directional reasoning" account, the learning rate will be higher when sources have neutral as opposed to partisan identities.

Experiment 1

In this experiment, we investigated the presence of evidence-based updating of trust towards partisan sources by disentangling the sources' partisanship (i.e., ingroup vs outgroup) from the sources' actual reliability (reliable vs unreliable). The experiment consisted of 40 rounds. In each round, participants performed a neutral estimation task and received a message from one of four different advisors: two ingroup advisors and two outgroup advisors. One ingroup and one outgroup advisor were fully reliable as they always provided accurate advice, while the other ingroup and outgroup advisor were fully unreliable as they always provided inaccurate advice (i.e., 2X2 within-subject design). At the end of each round, participants received feedback about the correct answer which also served as fact-checking information regarding the advisors' reliability. The presence and strength of feedback was manipulated across three between-subject conditions: (a) a no feedback condition in which feedback was entirely absent and thus learning was not possible, (b) a noisy feedback condition in which feedback was present but it was accurate 80% of the times, (c) a strong feedback condition in which feedback was present and always accurate. In the presence of strong feedback, we compared the learning rate when the advisors have partisan identities with a control condition in which the advisors have neutral identities.

Method

Participants

We recruited participants from the US through the online platform Academic Prolific. We selected participants who identified strongly or moderately with the Democratic or Republican party and voted for this party in the last two presidential elections in 2016 and 2020. We planned to recruit 560 participants (140 per cell) which would provide us .80 power to detect medium effect size (η^2

= .06) for main effects and interaction effects at the error probability of α = .017 (correction for testing three hypotheses, H1-H3).

In total, 621 participants completed the study but 17 participants were excluded as they consistently reported the same prior and posterior estimation across rounds. The final sample of 604 participants (M_{Age} = 44.1, SD_{Age} = 13.7) included 287 males and 317 females. 311 participants identified with the Democratic party while 293 participants identified with the Republican party. Participants received a fixed payment of $3 for completing the roughly 30-minute study and earned an additional bonus (average bonus was $1.7) based on the accuracy of their estimations (see below).

Belief Updating Task

The experiment consisted of 40 rounds. Figure 1 presents the structure of a typical round. In each round, participants were briefly (2 sec) presented with a picture that consists of green and orange pixels (100000 pixels in total). Participants were truthfully informed that in half of the pictures 51% of the pixels are green and 49% of the pixels are orange. In the other half of the pictures, 49% of the pixels are green and 51% of the pixels are orange. Participants were asked to estimate "How likely it is that the green (or orange) pixels are the majority" on a 0-100 scale. Following their first estimation (prior belief), participants received a message regarding the majority color from one of the four advisors.

The advisors had participated in previous sessions of the experiment where they had seen the same pictures that participants saw, and they had provided information about the majority color (i.e., advice) on a 0-100 scale. The message was presented in a binary form which consisted of the color that advisors identified as the majority. The message was presented for 3 seconds together with an avatar that contains the logo of the political party (Democratic or Republican party) with

which the advisor identified. Given that participants also identified with one of the two political parties, advisors were perceived as ingroup or outgroup members.

Upon receiving the message, participants provided a second estimate of the likelihood that green (or orange) is the majority color (i.e., posterior belief). At the end of each round, participants received feedback about the actual majority color and a reminder of the advisor's choice of color in this round. Combining these two pieces of information, participants can infer the veracity of the advisor's message. Across rounds, participants with accuracy motives should use this feedback to update their trust in each advisor. Importantly, participants received a bonus payment of max. $3 that was based on the accuracy of their first or second estimation in a randomly chosen round.

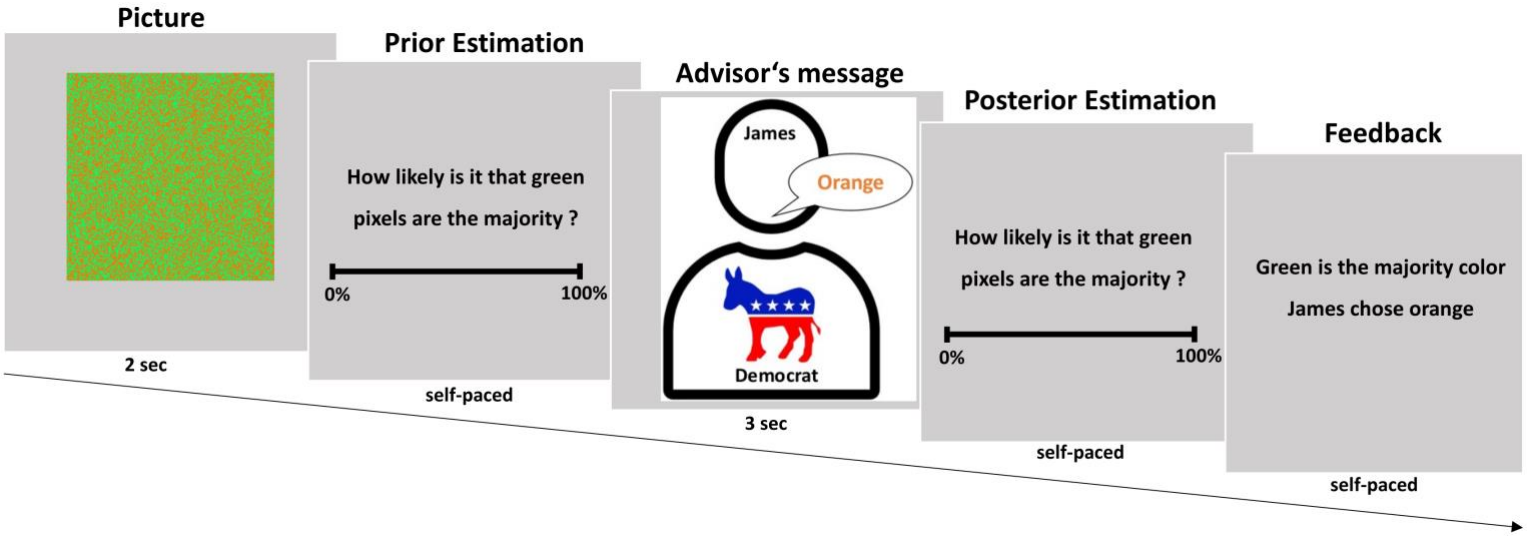

Figure 1. Structure of a typical round. Following the brief presentation of a picture with green and orange pixels, participants make a first estimation (prior) of the likelihood that green (or orange) is the majority color. Upon receiving a message from one of the four advisors, they make a second estimation (posterior). At the end of each round, they receive feedback about the correct answer and a reminder of advisor's message which serves as fact-checking information.

To disentangle the advisors' partisan identity from the advisors' reliability, we selected two advisors, one ingroup and one outgroup, who consistently gave accurate messages (10 out of 10

rounds) and two advisors, one ingroup and one outgroup, who consistently gave inaccurate messages (10 out of 10 rounds). Across rounds, we varied the advisor's partisan identity (ingroup vs outgroup), the advisor's reliability (reliable vs unreliable), the message (green or orange), and the actual majority color of the picture. For each of these characteristics, each of the two possible options was presented twenty times. Across the 40 rounds, these characteristics were combined in such ways that in half of the rounds, the message is accurate and in the other half it is inaccurate. Moreover, each of the four advisors chose green and orange as majority color with equal frequency to avoid a color bias. The 40 rounds were divided into five blocks of 8 rounds. Each possible combination of characteristics was presented with equal frequency across the five blocks but in a randomized order. The order was counterbalanced across participants.

In the aforementioned design, feedback about the correct answer was present in each round and was always accurate which constitutes the strong feedback condition. To examine how the presence and the strength of feedback affects trust towards the four advisors, we implemented two additional conditions, a "no feedback" condition where participants do not receive feedback at the end of each round and a "noisy feedback" condition where participants know in advance that feedback is correct 80% of the times. Moreover, to examine whether and to what extent advisors' partisan identities undermine evidence-based learning, we introduced a control condition where advisors have neutral identities by labelling them as members of the red or blue team.

Taken together, our design includes advisors' identity and advisors' reliability as the main within-subject independent variables. Given the presence of multiple rounds, block is an additional within-subject independent variable. Participants were randomly assigned to advisors' partisan or neutral identity conditions and to the three feedback conditions (i.e., strong, noisy, or no feedback). Advisors' neutral identity was combined with strong feedback as we aimed to establish

a baseline learning rate in the absence of partisan identities and when feedback is always correct (see Table 1).

Table 1. Conditions and sample size per cell

Between-subject variables	Advisors' Partisan Identity	Advisors' Neutral Identity
Strong Feedback	n = 140	n = 181
Noisy Feedback	n = 141	-
No Feedback	n = 142	-

Measure of affective polarization

Affective polarization is defined as the tendency to view rival social or political groups negatively and ingroup members positively (Green et al., 2002). In the last two decades, many studies have documented growing affective polarization that characterizes US politics in that Democrats and Republicans increasingly dislike and distrust each other (Druckman et al., 2021; Iyengar et al., 2019). To examine whether affective polarization predicts the early partisan bias in trust and the learning rate, we elicited participants' feelings (positive or negative) towards the two partisan groups. In line with previous work (Wagner, 2021), we used the difference between ingroup and outgroup feelings as an index of affective polarization. We also explored whether feelings towards ingroup and outgroup independently predict the early partisan bias in trust and the learning rate (see Supplemental Material).

Belief Updating

To measure the extent to which participants updated their beliefs in response to an advisor's message, we used the Log Likelihood Ratio (LLR). The LLR quantifies the amount of information that participants incorporate in their belief upon receiving the advisor's message. Let us assume

that participant's prior probability that the majority color is green is P(G). Then, participant's prior probability that the majority color is orange is P(O) = 1 – P(G). According to Bayes rule, the posterior probability that the majority color is green after receiving the message (A) is:

Equation 1:

$$P(G|A) = \frac{P(G) \cdot P(A|G)}{P(G) \cdot P(A|G) + P(O) \cdot P(A|O)}$$

where P(A|G) is the probability of receiving the message A given the majority color is green and P(A|O) is the probability of receiving the same message given the majority color is orange.

The LLR of the received message is the difference between the log-posterior odds and the log-prior odds:

Equation 2:

$$log\left(\frac{P(A|G)}{P(A|O)}\right) = log\left(\frac{P(G|A)}{P(O|A)}\right) - log\left(\frac{P(G)}{P(O)}\right)$$

Note that the right-hand side in Equation 2 corresponds to the reported prior and posterior beliefs and therefore the LLR can be computed from participants' estimations. Using the LLR rather than the absolute difference between posterior and prior estimation allows us to account for the strength of prior beliefs. For instance, the LLR that we obtain when beliefs are updated from 80% to 90% is much larger than the LLR we obtain when beliefs are updated from 50% to 60%. Although in both cases the absolute belief change is the same in percentage points, in the former case the participant has incorporated much more information than in the latter case. In other words, the participant has interpreted the message as much stronger evidence in the former than in the latter case. We assigned a positive value to LLR when the direction of belief updating is consistent with the message participants received and negative value when the direction of belief updating

contradicts the received message. In this respect, a higher LLR in response to ingroup compared to outgroup messages is an indication of partisan bias in trust.

Statistical analysis

To test our hypotheses, we used linear mixed effects models which offer several advantages over classic ANOVA when observations are nested within subjects (Barr et al., 2013; Brauer & Curtin, 2018). We sequentially fitted different linear mixed effect models to the round-level data to determine whether a fixed effect model, a random intercept model, or a random intercept and random slope model provides the best model fit. We used a Likelihood Ratio Test to statistically evaluate goodness of fit. For all statistical tests described below, the most complex random-effects structure available provided the best fit to the data.

We first examined whether participants exhibited an early partisan bias in trust towards the four advisors (H1). To test the presence of an early partisan bias, we focused only on conditions in which advisors have partisan identities and we limited the analysis to the first time point (i.e., first block out of the five blocks of rounds). We fitted a linear mixed effect model using advisors' identity (ingroup vs outgroup) and advisors' reliability (reliable vs unreliable) as within-subject independent variables, presence and strength of feedback as a between-subject independent variable, and LLR as the dependent variable.

We then examined whether the partisan bias in trust persisted over time or whether participants gradually updated their trust in the presence of feedback (H2). Here, we focused only on conditions in which advisors' identity is partisan and feedback is present. We fitted a linear mixed effect model using advisors' identity, advisors' reliability, and time as within-subject independent variables, strength of feedback as between-subject independent variable, and LLR as the dependent variable.

Lastly, we examined whether the evidence-based updating of trust toward reliable and unreliable sources differs depending on whether advisors' identities are partisan or neutral (H3). Here, we focused only on conditions in which participants received strong feedback. We fitted a linear mixed effect model using advisors' identity, advisors' reliability, and time as within-subject independent variables, presence of partisan identities as between-subject independent variable, and LLR as the dependent variable.

Significant interaction effects were further explored by testing the effect of one variable separately in each level of the other. In an exploratory analysis, we tested whether participants' identity (Democrats vs Republicans) moderated the aforementioned effects.

Results

We observed a positive average LLR which indicates that participants overall trusted the advisors irrespective of partisanship or actual reliability (Figure 2).

Early partisan bias in trust

Consistent with H1a, supporters of both political parties exhibited an early partisan bias ($F(1, 1110) = 14.29$, $p < .001$, $\eta_p^2 = .010$) by incorporating information from ingroup sources more than outgroup sources, irrespective of advisors' actual reliability and the strength of feedback (Block 1 in Figure 2a). To further examine the nature of this effect, we regressed individual differences in early partisan bias against an index of affective polarization. This analysis revealed a significant positive correlation ($b = .006$, $t(371) = 2.92$, $p = .004$) indicating that the more polarized partisans' feelings towards ingroup and outgroup are, the stronger early partisan bias they displayed.

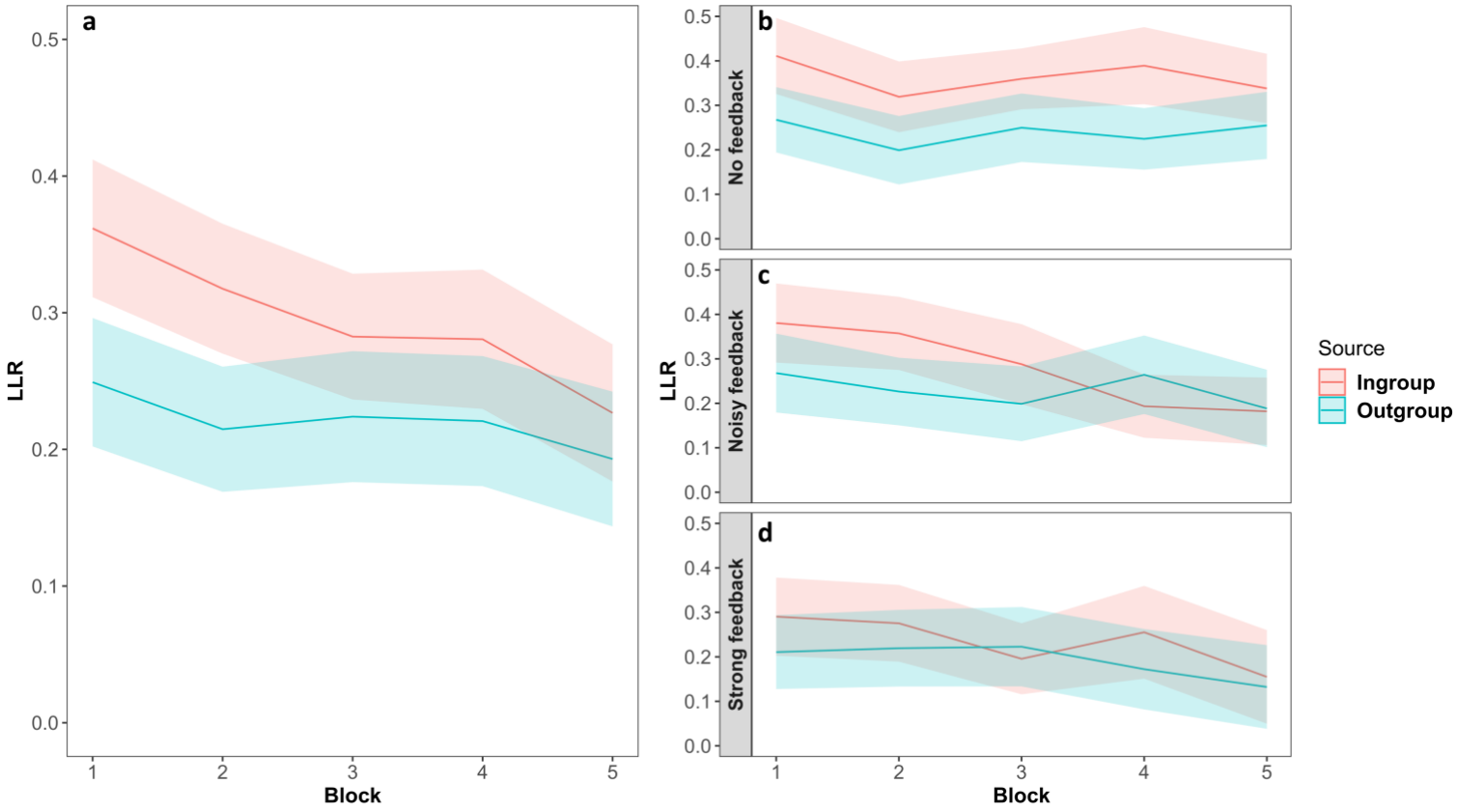

Figure 2. Time evolution of average LLR indicating trust towards ingroup (red) and outgroup (blue) partisan sources for all feedback conditions (a) and separately for the no feedback (b), the noisy feedback (c), and the strong feedback (d) condition. Bands around the averages indicate the 95% confidence interval of the mean.

Persistence of partisan bias and evidence-based updating of trust

As expected, in the absence of feedback, the partisan bias in trust persisted throughout the experiment ($F(1, 2298) = 18.65$, $p < .001$, $\eta_p^2 = .009$) (Figure 2a) and participants did not distinguish between reliable and unreliable sources ($F(1, 437) = 0.20$, $p = .655$) (Figure 3a).

In contrast, in the presence of strong or noisy feedback, we observed a decline of partisan bias over time ($F(1, 4155) = 5.36$, $p = .021$, $\eta_p^2 = .001$) which was mainly driven by a declining trust towards ingroup sources ($F(1, 243) = 22.90$, $p < .001$, $\eta_p^2 = .090$) (Figure 2b-c). Post-hoc tests revealed that the partisan bias in trust remained significant only in the first two blocks. Notably,

participants of both political groups gradually learned to trust reliable sources more than unreliable sources when feedback was present ($F(1, 429) = 21.13$, $p < .001$, $\eta_p^2 = .050$) (Figure 3b-c). This gradual learning effect was mainly driven by a declining trust towards unreliable sources ($F(1, 243) = 40.53$, $p < .001$, $\eta_p^2 = .140$). Taken together, the declining partisan bias and the evidence-based updating of trust are consistent with H2b and with the "accuracy motives with biased priors" account (see Supplemental Material for more detailed analysis).

Contrary to H2c, the strength of feedback did not have an impact on the decline of the partisan bias ($F(1, 4155) = 2.41$, $p = .121$) and the evidence-based updating of trust ($F(1, 429) = 1.13$, $p = .288$). Specifically, partisans learned to trust reliable sources more than unreliable sources in the face of both noisy ($F(1, 259) = 8.70$, $p = .003$, $\eta_p^2 = .030$) and strong ($F(1, 192) = 12.36$, $p < .001$, $\eta_p^2 = .060$) feedback. The aforementioned effects were present in both Democrats and Republicans. Moreover, affective polarization ($b = .005$, $t(371) = 1.28$, $p = .200$) did not predict the extent to which partisans learned to trust reliable over unreliable advisors.

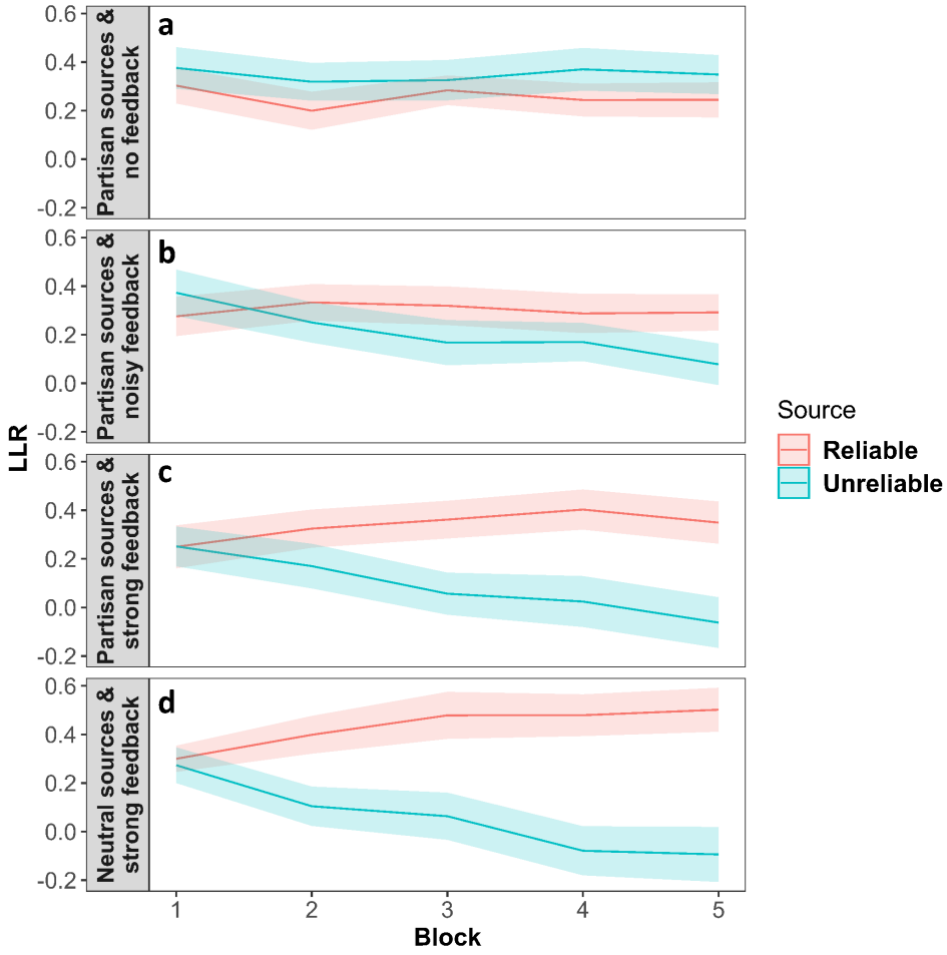

Figure 3. Time evolution of average LLR indicating trust towards reliable (red) and unreliable (blue) sources separately for partisan sources and no feedback (a), partisan sources and noisy feedback (b), partisan sources and strong feedback (c), and neutral sources and strong feedback (d). Bands around the averages indicate the 95% confidence interval of the mean.

Evidence-based updating of trust towards partisan vs neutral sources

Consistent with H3a and the "accuracy motives with biased priors" account, partisan sources, compared to neutral sources, did not slow down the learning process ($F(1, 416) = 0.61$, $p = .435$) (Figure 3c-d). For both partisan and neutral sources, post-hoc tests confirmed that trust towards the unreliable sources declined over time ($F(1, 279) = 52.27$, $p < .001$, $\eta_p^2 = .160$), while the reliable sources were trusted more over time ($F(1, 279) = 16.27$, $p = .001$, $\eta_p^2 = .060$).

Discussion

Both Democrats and Republicans initially incorporated information from ingroup sources more than outgroup sources and the extent of this partisan bias was positively associated with affective polarization and positive feelings towards ingroup. In the presence of strong or noisy feedback, the partisan bias gradually disappeared as Democrats and Republicans gradually trusted less the unreliable sources. The extent of learning correlated with negative feelings towards the outgroup. Importantly, this learning process did not differ from a control condition in which sources had neutral rather than partisan identities suggesting that sources' partisanship did not undermine learning. Taken together, these findings are consistent with the "accuracy motives with biased priors" account in which partisans initially hold biased credibility impressions about partisan sources but these impressions are gradually updated based on feedback.

Experiment 2

In Experiment 1, the different sources consistently provided either accurate (i.e., reliable sources) or inaccurate (unreliable sources) information. However, sources that communicate political information are rarely characterized by this absolute level of (in)accuracy. Credible sources may in some occasions provide inaccurate information, while dubious sources occasionally provide accurate information. To emulate this noisy environment, we conducted a second preregistered experiment in which partisan sources share a mixture of accurate and inaccurate information and thus there is ambiguity about the actual source reliability. To facilitate the learning process in this noisy environment, participants received information only from two sources, an ingroup and an outgroup source.

In Experiment 1, ingroup and outgroup sources were equally reliable (i.e., one reliable and one unreliable source from each partisan group). In contrast, in this experiment, the ingroup source provides most of the time inaccurate information while the outgroup source is most of the time accurate. In this setting, accuracy motives are in direct conflict with identity-protective motives and thus any decline of partisan bias in trust reflects evidence-based updating of trust. We also introduced a "weak evidence" condition in which both the ingroup and outgroup source provide accurate information in only half of the rounds. This control condition allows us to establish a baseline level of partisan bias in a highly noisy environment.

In Experiment 1, the content of the information was neutral (i.e., majority color in the picture). In this experiment, we introduced an additional condition to examine whether political information would affect the rate with which people update their trust towards partisan sources. Similar to Experiment 1, we introduced a control condition with neutral source identities that allowed us to quantify the baseline learning rate when the partisan bias is absent.

Method

All manipulations, measures, hypotheses, data analysis, sample size, and data exclusions of the study were preregistered prior to data collection (https://osf.io/mtfy2/?view_only=4b49d393c56040eb80f3b123b374c77d). The experiments were programmed in Qualtrics.

Participants

We recruited participants from the US through the online platform Academic Prolific. As reported in the preregistered protocol, we sought to recruit 720 participants (120 per cell) which would provide us .80 power to detect medium effect size ($\eta^2 = .06$) for main effects and interaction effects at the error probability of $\alpha=.017$ (correction for testing three preregistered hypotheses). In total, 754 participants completed the study but 20 participants did not fulfill our preregistered inclusion criteria (consistently reported the same prior and posterior estimation across rounds). The final sample of 734 participants ($M_{Age} = 46.9$, $SD_{Age} = 13.7$) included 358 males and 376 females. 379 participants identified strongly or moderately with the Democratic party while 355 participants identified strongly or moderately with the Republican party. Similar to Experiment 1, participants received a fixed payment of $3 for completing the roughly 30-minute study and an additional bonus of max. $3 that was based on the accuracy of participants' prior or posterior estimations in a randomly chosen round (average bonus was $1,6).

Belief Updating Task

We modified the Belief Updating task of Experiment 1 in that partisan sources provided a mixture of accurate and inaccurate information (i.e., noisy environment). The experiment consisted of 36 rounds and participants received messages from an ingroup and an outgroup advisor. Participants performed the same estimation task (i.e., majority color), but the pictures consisted of red and blue pixels rather than green and orange pixels. In each round, participants received accurate

feedback about the actual majority color and a reminder of the advisor's choice of color in this round which allowed inferences about the advisors' actual credibility.

To emulate the noise that characterizes real-life settings, the outgroup, reliable advisor sent accurate messages in 15 out of 18 rounds and the ingroup, unreliable advisor sent inaccurate messages in 15 out of 18 rounds. To avoid any order effects, inaccurate messages from the reliable advisor and accurate messages from the unreliable advisor were always presented in the 3rd, 9th, and 15th round. Each round differed with respect to the advisor's identity, the message (red or blue), and the actual majority color of the picture. The 36 rounds were divided into three blocks of 12 rounds. Each possible combination of characteristics was presented with equal frequency across the three blocks but in a randomized order. The order was counterbalanced across participants. Across the 36 rounds, these three characteristics were combined such that messages are accurate in half of the rounds. To avoid a color bias, each advisor chose blue and red as the majority color with equal frequency. We also introduced a "weak evidence" condition where both ingroup and outgroup advisors sent accurate messages in half of the rounds and thus there was no clear evidence that one of the two advisors is more reliable than the other.

We also introduced a political context, by using the contentious issue of the winner of the 2020 US presidential election which still polarizes the supporters of the two political parties (Botvinik-Nezer et al., 2023; Kahn, 2021). Specifically, we selected some counties throughout the US where, according to the official results by the Federal Election Commission, the winning party won by a narrow margin (i.e., roughly 51% of the eligible votes). We visually represented the electoral result in these "narrow margin" counties with pictures that consist of 100000 blue and red pixels. Blue pixels represent the percent of votes for the Democratic party and red pixels represent the percent of votes for the Republican party. In essence, participants performed the same visual estimation task using the same stimuli as in the neutral information condition.

In this political context, participants were asked to estimate "How likely it is that Democrats (or Republicans) won in this county". Similar to the neutral content, the message came from other participants who took part in this study. The message was presented in a binary form and consisted of the political party that, according to the advisor, won the elections in this county (again 50% was used as a threshold to binarize the advisors' response). The political content introduces a desirability bias in that the desired belief for Democrats and Republicans is that their party won (Dahlke & Hancock, 2022). To ensure that the desirability bias does not confound the partisan bias, both ingroup and outgroup advisors sent desirable messages (i.e., own political party won elections) in half of the rounds and undesirable messages (i.e., rival political party won elections) in the other half of the rounds.

Similar to Experiment 1, we introduced a control condition where both advisors have neutral identities (i.e., advisor A vs advisor B). Our design includes advisor's identity/reliability and time (i.e., three blocks) as the main within-subject independent variables. Moreover, our design includes the following between-subject independent variables: content (neutral vs political), evidence about the advisors' reliability (strong vs weak), and advisors' identities (partisan vs neutral identity). These three between-subject variables are not fully orthogonal to each other (see Table 2), as we did not combine advisors' neutral identity (i.e., no partisan bias) with weak evidence about the advisors' reliability (i.e., no learning).

Table 2. Conditions and sample size per cell

Between-subject variables	**Advisors' Partisan Identity**		**Advisors' Neutral Identity**
	Strong Evidence about source reliability	**Weak Evidence about source reliability**	**Strong Evidence about source reliability**
Neutral	n = 118	n = 117	n = 159

Message Content			
Political	n = 121	n = 121	n = 118
Message Content			

Additional measures

We measured reliability impressions by eliciting participants' beliefs about the reliability of advisors' responses during the task (i.e., frequency of correct messages sent by ingroup or outgroup advisor). These reliability impressions were elicited at the beginning (prior) and at the end (posterior) of the experimental session. Similar to Experiment 1, we also elicited participants' feelings (positive or negative) towards the two groups, and we used the difference between ingroup and outgroup feelings as a measure of affective polarization. Lastly, a short memory task (recognition task) was administered after the belief updating task to control for differences in memory skills.

Belief Updating

To measure the extent to which participants update their beliefs in response to an advisor's message, we used the LLR which was computed as in Experiment 1.

Statistical analysis

As in Experiment 1, we sequentially fitted different linear mixed effect models to the round-level data to determine whether a fixed effect model, a random intercept model, or a random intercept and random slope model provides the best model fit. For all statistical tests described below, the most complex random-effects structure available provided the best fit to the data.

To test H1 (early partisan bias), we focused only on the condition in which advisors had partisan identities and we limited the analysis to the early phase (i.e., first block) of the experiment. We

fitted a linear mixed effect model using advisors' identity as within-subject independent variable and strength of evidence as well as message content as between-subject independent variables.

To test hypothesis H2 (persistence of partisan bias and evidence-based updating of trust), we focused only on the condition in which advisors had partisan identities. We fitted a linear mixed effect model using advisors' identity and time as within-subject independent variables. Strength of evidence and message content were added to the model as between-subject independent variables.

To test hypothesis H3 (differences in the learning process between partisan and neutral identities), we focused only on the strong evidence condition where advisors are either clearly reliable or unreliable. We fitted a linear mixed effect model using advisors' reliability and time as within-subject independent variables and presence of partisan identities as between-subject independent variable. In all aforementioned models, the LLR was used as dependent variable.

Results

There was a positive average LLR indicating that participants overall trusted the two partisan sources throughout the experiment.

Early partisan bias in trust

Consistent with H1a, Democrats and Republicans exhibited an early partisan bias ($F(1, 465) = 6.54$, $p = .011$, $\eta_p^2 = .010$) in that they incorporated information from the ingroup source more than the outgroup source in the first block of rounds (Block 1 in Figure 4a). Contrary to both H1b and H1c, the content of the information (i.e., neutral vs political) had no impact on the early partisan bias. Furthermore, the early partisan bias was the same in the presence of both strong and weak evidence about the actual source reliability. Similar to Experiment 1, individual differences in affective polarization ($b = .002$, $t(465) = 3.47$, $p < .001$) positively predicted the early partisan bias.

As expected, the partisan bias was also present in self-reported reliability impressions at the beginning of the experiment ($t(461) = 14.21$, $p < .001$). However, the self-reported reliability impressions did not predict the behavioral measure of partisan bias ($b = .004$, $t(460) = 1.07$, $p = .285$).

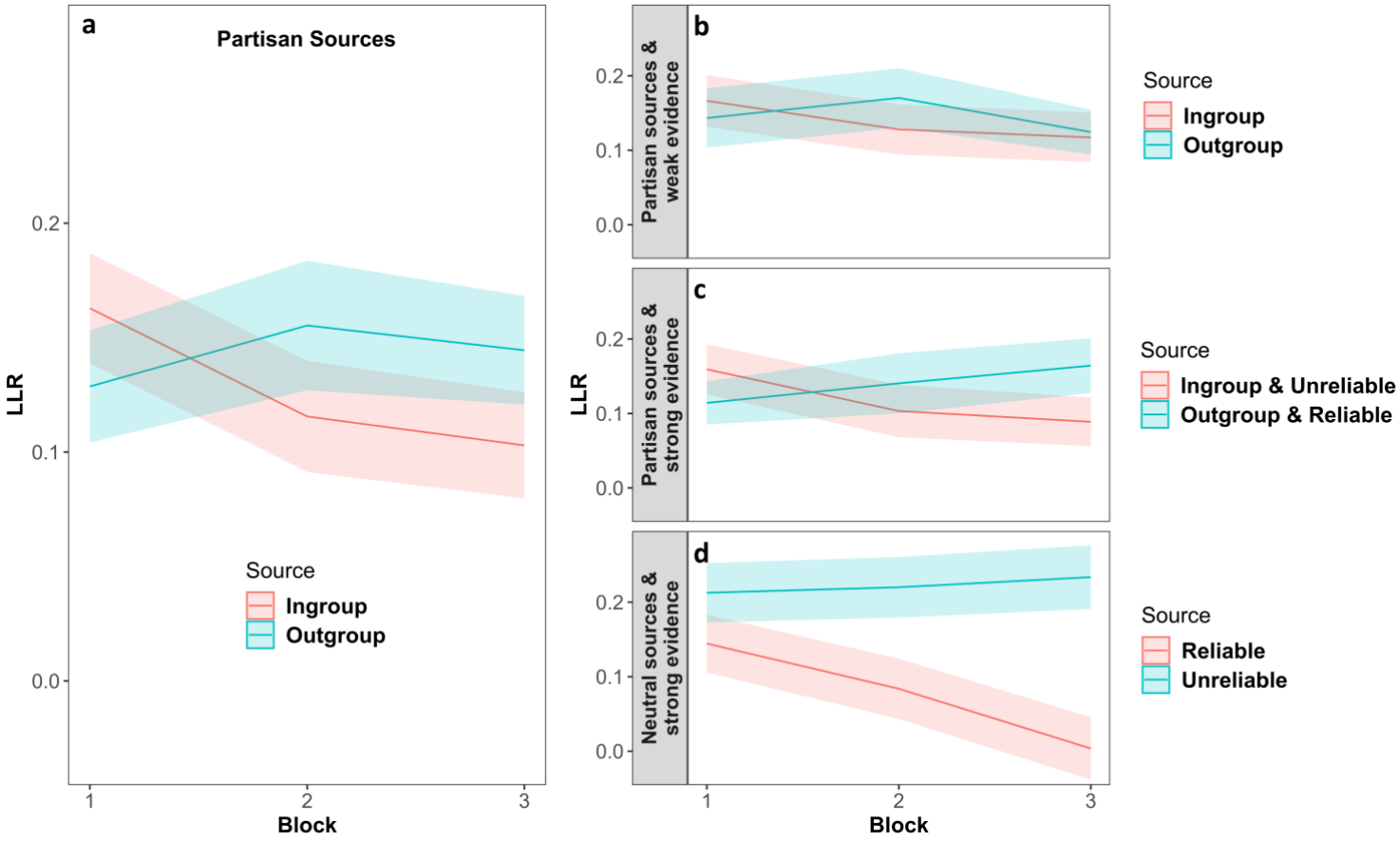

Figure 4. Time evolution of average log likelihood ratio (LLR) indicating trust towards the ingroup and unreliable source (red) and the outgroup and reliable (blue) source for all conditions that included partisan sources (a) and separately for the condition that combined partisan sources with weak evidence (b), partisan sources with strong evidence (c), and neutral sources with strong evidence. Bands around the averages indicate the 95% confidence interval of the mean.

Persistence of partisan bias and evidence-based updating of trust

Consistent with H2b and the "accuracy motives with biased priors" account, the partisan bias gradually declined and even changed direction (Figure 4a), as both Democrats and Republicans gradually updated their trust towards partisan sources based on evidence ($F(1, 1584) = 13.22$, $p < .001$, $\eta_p^2 = .007$). This learning effect was mainly driven by the gradually declining trust towards the ingroup and unreliable advisor ($F(1, 933) = 19.45$, $p < .001$, $\eta_p^2 = .020$), while trust towards the outgroup and reliable advisor did not change over time ($F(1, 933) = 1.25$, $p = .263$).

Consistent with H2c, the learning rate was lower in the face of weak compared to strong evidence ($F(1, 1584) = 4.76$, $p = .029$, $\eta_p^2 = .003$) suggesting that a noisy environment impedes the learning process (Figure 4b-c). In fact, post-hoc tests revealed that the partisan bias was significantly reduced in the presence of strong evidence ($F(1, 614) = 14.65$, $p < .001$, $\eta_p^2 = .020$) but not in the presence of weak evidence ($F(1, 924) = 1.16$, $p = .282$, $\eta_p^2 = .001$).

In contrast to H2d, the learning process did not differ between neutral and political content ($F(1, 1587) = 2.20$, $p = .138$). We also examined the learning process separately in the two types of content. This exploratory analysis revealed that the learning effect was present when the content was neutral ($F(1, 763) = 11.48$, $p < .001$, $\eta_p^2 = .010$) but it was absent when the content was political ($F(1, 924) = 2.82$, $p = .093$).

Notably, affective polarization did not predict the evidence-based updating of trust towards the two advisors indicating that the learning process is independent from the degree of affective polarization. In line with the behavioral measure, the partisan bias in self-reported reliability impressions was significantly reduced throughout the experiment ($t(461) = -9.31$, $p < .001$) and was absent at the end of the experiment ($t(461) = 0.54$, $p = .589$). However, the decline of the self-reported partisan bias did not predict the gradual decline of the behavioral partisan bias ($b = .006$, $t(460) = 1.50$, $p = .134$).

Evidence-based updating of trust towards partisan vs neutral sources

Consistent with H3a and the «accuracy motives with biased priors" account, the presence of partisan (vs neutral) identities had no impact on how fast Democrats and Republicans learned to trust reliable sources ($F(1, 1064) = 0.57$, $p = .451$) (Figure 5c-d). In fact, evidence-based updating of trust was substantial irrespective of whether advisors' identity was partisan or not ($F(1, 1064) = 26.59$, $p < .001$, $\eta_p^2 = .020$). Post-hoc tests showed that the unreliable advisor was trusted less over time ($F(1, 942) = 6.25$, $p = .013$, $\eta_p^2 = .006$), while the reliable advisor was trusted more over time ($F(1, 501) = 49.63$, $p < .001$, $\eta_p^2 = .090$).

Nonetheless, throughout the experiment the reliability effect was more pronounced for neutral compared to partisan source identities ($F(1, 2004) = 14.79$, $p < .001$, $\eta_p^2 = .007$). Specifically, for neutral source identities, we observed a clear preference for information coming from the reliable over the unreliable source ($F(1, 1333) = 91.23$, $p < .001$, $\eta_p^2 = .060$). In contrast, for partisan source identities, there was no clear overall preference for the reliable source ($F(1, 943) = 3.53$, $p = .060$, $\eta_p^2 = .004$) as participants initially trusted the ingroup unreliable source more than the outgroup reliable source. Nonetheless, the learning rate did not differ between neutral and partisan source identities suggesting that this difference can be explained by the early partisan bias which was present only for partisan identities (see Figure 4c-d).

Discussion

Similar to Experiment 1, Democrats and Republicans initially incorporated information from the ingroup source more than the outgroup source irrespective of the content (neutral or political). Affective polarization and feelings towards the ingroup predicted the extent of this early partisan bias. In the face of strong evidence, Democrats and Republicans gradually trusted more the outgroup source and gradually trusted less the ingroup source. In the presence of weak evidence (i.e., ingroup and outgroup advisors are equally reliable), partisans gradually trusted less the

ingroup source but trust towards the outgroup source did not change significantly. Although the learning process did not differ significantly across neutral and political context, a separate analysis for the two contexts revealed that the learning process was significant only in the neutral context. Importantly, this learning process was the same for neutral and partisan sources suggesting that sources' partisanship does not activate identity-protective motives. However, throughout the experiment, the reliability effect on trust was more pronounced in the presence of neutral sources. This difference is explained by the early partisan bias which has a lasting impact on trust levels throughout the experiment.

General Discussion

Cognitive scientists have proposed two competing mechanisms for the emergence and persistence of erroneous and polarized political beliefs: (1) partisans engage in directional reasoning and distort their inference process to protect their political identities and affirm their ideology, or (2) they are motivated by accuracy but use partisanship as a heuristic to identify reliable sources of information. Using a dynamic experimental setting that allows partisans to gradually discern the credible from the dubious partisan sources, we provide substantial evidence in favor of the second mechanism. Both Democrats and Republicans initially displayed a partisan bias in that they incorporated information from ingroup sources more than outgroup sources. However, this partisan bias gradually disappeared or even changed direction after exposure to evidence that outgroup sources are equally or more reliable than ingroup sources, respectively. This learning effect was present when partisan sources shared either neutral or political information. However, an exploratory analysis suggests that the learning effect is more pronounced in the face of neutral compared to political information. Moreover, the learning process was not sensitive to the strength of external feedback but was negatively affected by the degree of uncertainty regarding actual source reliability. Compared to neutral source identities, the presence of partisan source identities did not undermine the learning process further supporting the view that partisans are driven by accuracy rather than identity-protective motives.

The early partisan bias regarding trust in information sources is consistent with an extensive body of literature that has documented partisan biases in various behavioral measures and contexts (Baron & Jost, 2019; Carlin & Love, 2018; Ditto et al., 2019; Leeper & Slothuus, 2014). Nonetheless, the nature of this partisan bias remains relatively elusive. One plausible interpretation is that this bias is driven by beliefs about the credibility and trustworthiness of partisan groups which are based on stereotypes (Dovidio et al., 2010; Fiske et al., 2002) or selective exposure to information

(Derreumaux et al., 2022; Levendusky, 2013; Mothes & Ohme, 2019). In line with this interpretation, self-reported reliability ratings showed a similar pattern of an initial partisan bias that disappeared after exposure to feedback. However, the early partisan bias and the degree of learning observed at the behavioral level did not correlate with self-reported reliability ratings. This discrepancy between behavioral and self-reported measures is reported in other types of biases as well. For instance, the racial bias in trust during economic transactions was associated with implicit rather than explicit racial bias (Stanley et al., 2011).

An alternative interpretation is that the biased incorporation of information from partisan sources reflects an implicit bias that is driven by affective prejudice towards the two groups (Amodio et al., 2003; Cuddy et al., 2009). Consistent with this interpretation, in both experiments, polarized feelings towards the two partisan groups predicted the early behavioral bias but not the degree of evidence-based updating of trust. Taken together these findings suggest that, in the absence of other information, partisans rely on an affective implicit bias to solve the problem of which source of information to trust. However, once external feedback becomes available, partisans start basing their decisions on concrete evidence about source credibility (Schulz et al., 2023).

In both experiments, the partisan bias gradually declined after exposure to external feedback and was even reversed when the outgroup sources were more reliable than the ingroup sources. Consistent with Bayesian reasoning, this learning process was sensitive to the degree of uncertainty regarding actual source reliability. On the other hand, the introduction of noise in the external feedback had no impact on the learning process indicating that partisans did not discriminate between perfectly accurate (i.e., 100%) and highly informative (i.e., 80%) external feedback. This strategy is optimal in the present setting given the absence of any other information about source reliability.

In both experiments, the evidence-based updating of trust was predominantly driven by a gradual decline of trust towards unreliable sources rather than an increasing trust towards reliable sources. One plausible explanation for this heightened sensitivity to negative feedback is that people initially consider any source of information as potentially helpful (Schulz et al., 2023) and thus are more likely to generate large prediction errors in the face of negative rather than positive feedback. Indeed, in both experiments, partisans initially incorporated information from all sources irrespective of their identity. Taken together, the incorporation of both strong and noisy external feedback, the sensitivity to the strength of evidence regarding actual source reliability, and the increased susceptibility to negative feedback suggest that people follow Bayesian learning dynamics. This pattern is consistent with previous studies reporting evidence-based updating of prior impressions about others' moral character (M. J. Kim et al., 2021; Mende-Siedlecki, 2018), trustworthiness in economic transactions (Traast et al., 2023) and credibility as sources of information (Schulz et al., 2023).

Previous work has shown that contexts that activate partisan identities, such as polarizing political issues, activate identity-protective or ideology-affirming motives (Grace et al., 2008; Hogg et al., 1995; Tajfel & Turner, 2004) and thus enhance directional reasoning. However, our results do not support this hypothesis as the early partisan bias in trust did not differ across neutral and political contexts suggesting that the partisan bias is driven by the mere presence of partisan source identities rather than the polarizing content of the information. Moreover, the evidence-based updating of trust did not differ across the two contexts indicating that partisans are motivated by accuracy in both neutral and political contexts. However, this conclusion warrants further investigation as a separate analysis revealed that the learning effect is present only in the face of neutral information.

Despite the polarizing nature of the topic (i.e., fraud in the 2020 US presidential election) (Botvinik-Nezer et al., 2023; Kahn, 2021), we did not observe a desirability bias in that there was no asymmetric incorporation of desirable and undesirable information from partisan sources. However, partisans expressed a strong desirability bias during the elicitation of prior beliefs. This effect was mainly driven by Democrats reporting that the Democratic party won the elections (see Supplemental Material). This pattern further confirms the view that partisans hold biased prior beliefs on this polarizing issue but have accuracy motives and are willing to incorporate new information even if it comes from outgroup sources and is undesirable.

Given that partisan sources often share politically concordant information, people develop second-order beliefs about the biases of partisan groups (Bogart & Lees, 2023; Mernyk et al., 2022). When this is the case, any mismatch between information desirability and source identity is unexpected and can be perceived as more informative. Therefore, desirable information may be incorporated to a larger extent when it comes from outgroup sources and vice versa for undesirable information. In line with this argument, we found an interaction between source identity and information desirability in that the early partisan bias was present only for undesirable information (see Supplemental Material). This finding is consistent with evidence that partisans are characterized by negative meta-perceptions about outgroups (Bogart & Lees, 2023; Mernyk et al., 2022) which lead to rejection of undesirable information when it comes from rival partisan groups. However, our study design does not allow us to draw firm conclusions on the potential interactions between desirability and identity bias, yet our preliminary findings warrant further investigation especially about the role of meta-perceptions in the learning process.

The evidence-based updating of trust indicates the presence of accuracy motives but does not exclude the possibility that identity-protective motives are simultaneously present and slow down the learning process. Furthermore, even accuracy-motivated individuals may maintain their strong

prior impressions about partisan sources in a procedurally rational way by generating auxiliary explanations for identity-incongruent feedback (Gershman, 2019; M. Kim et al., 2020). To test these two plausible explanations, we introduced a control condition in which sources have neutral identities and thus these two factors become irrelevant. In both experiments, the learning process did not differ between partisan and neutral sources suggesting that identity-protective motives and auxiliary explanations, if present, played a trivial role. Despite the similar learning rates across the two identity conditions, the early partisan bias acted as a handicap in that partisans required more time and a larger amount of evidence to successfully discern credible from dubious partisan sources (Derreumaux et al., 2022, 2023).

An ongoing debate pertains to the presence of ideological asymmetries in the processing of information (Allcott & Gentzkow, 2017; Guay & Johnston, 2022). Some studies showed that liberals and conservatives or Democrats and Republicans display important asymmetries in the way they process information (Baron & Jost, 2019; DeVerna et al., 2022; Van Der Linden et al., 2021), while other studies suggest that these processes are independent from ideology and moral values (Brandt et al., 2014; Ditto et al., 2019). Our findings are in line with the latter view as the two partisan groups displayed a similar pattern of an early partisan bias and subsequent evidence-based updating of trust.

It is difficult to reconcile our finding that people follow Bayesian learning dynamics with the growing trend of belief polarization and resistance to evidence (Allcott & Gentzkow, 2017; Van Der Linden, 2022). Going back to our initial example, it is unclear why Donald Trump still enjoys the trust of a large proportion of Republicans despite the extensive debunking of his false and misleading claims (Baker, 2018; Swire et al., 2017; The Washington Post, 2021). One plausible explanation is that in highly polarized political contexts, accurate information is not abundant and seeking evidence that debunks misinformation is an active and effortful process (Lewandowsky et

al., 2012; Roozenbeek et al., 2023). Moreover, partisans selectively expose themselves to sources they deem as credible and rarely verify the claims of these sources (Derreumaux et al., 2022; Levendusky, 2013; Mothes & Ohme, 2019). This tendency has recently increased due to the declining trust in mainstream media and the emergence of alternative media sources that do not adhere to journalistic standards (Hameleers et al., 2022; Michailidou & Trenz, 2021). In contrast to this noisy environment, our setting exposes partisans to information from both sides and provides immediate and accurate external feedback. Nonetheless, even in this highly controlled setting, we found that uncertainty about actual source reliability plays a crucial role as it impedes learning and preserves the partisan bias in trust.

Our setting also provides external monetary incentives for accuracy which are not present in the real world. Recent experimental work has provided robust evidence that accuracy prompts (Pennycook & Rand, 2022; Roozenbeek et al., 2021) and accuracy incentives (Rathje et al., 2023) can mitigate misinformation susceptibility and resistance to evidence. Both motivational and cognitive mechanisms can account for this effect. The motivational account posits that accuracy prompts and incentives render accuracy motives more salient than identity-protective motives and thus curtail directional reasoning (Kahan, 2016; Van Bavel & Pereira, 2018).

The cognitive account posits that lack of reflective reasoning and limited attentional resources lead to intuitive and automatic processing of information (Bago et al., 2020; Pennycook & Rand, 2019, 2021). In this respect, accuracy prompts and incentives promote reflective reasoning which improves discernment between true and false information (Pennycook & Rand, 2019, 2022). The cognitive account may be particularly relevant in our learning paradigm in which people need to keep track of fact-checking information for multiple sources across many rounds. However, despite the huge cognitive load that our paradigm imposes on working memory, we found that individual

differences in working memory skills did not predict evidence-based updating of trust (see Supplemental Material).

Taken together, our results have important implications regarding strategies that aim to combat misinformation and belief polarization phenomena. The presence of an early partisan bias and the absence of directional reasoning suggest that interventions should focus on mitigating prior prejudice towards partisan groups or correcting biased prior beliefs about partisan credibility. However, future research should examine whether the absence of directional reasoning is an artifact of our relatively neutral setting, the presence of external accuracy incentives, and the immediate provision of reliable fact-checking information.

Constraints on Generality

Given the increased polarization that characterizes US politics and the fact that the US has been the overwhelming focus of research on polarization (Boxell et al., 2024; Iyengar et al., 2019), an outstanding question is whether the present findings and especially the early partisan bias in neutral experimental settings can be generalized to partisan groups in other countries (Wagner, 2021). Another unaddressed question is whether the evidence-based updating of trust has a long-lasting impact and can be generalized to other contexts. Previous work has shown that implicit and explicit impressions about others are often sticky (Hernandez & Minor, 2015; M. Kim et al., 2020; Stanley et al., 2011) and upon a change of context people rely heavily on initial trait representations (Hackel et al., 2020). In this respect, our intervention may be limited in that partisans will display the same bias again on novel targets, like other partisan sources or different topics, or even on familiar targets in future encounters (Traast et al., 2023). Contrary to this view, other evidence suggests that people use their previous experience with members of one group to adjust their behavior towards other individuals of the same group (Vermue et al., 2019).

Conclusion

Across two experiments, we provide a host of evidence that Democrats and Republicans initially use partisanship as a heuristic to identify credible sources of information but do not engage in directional reasoning. Instead, they incorporate external feedback and gradually learn to discern the credible from the dubious sources irrespective of partisanship. Our findings provide novel insights into the underlying cognitive and motivational mechanisms of belief and impression updating in political contexts and can contribute to the development of successful communication strategies that combat misinformation and belief polarization phenomena.

References

Abrams, D., Rutland, A., Cameron, L., & Marques, Josém. (2003). The development of subjective group dynamics: When in-group bias gets specific. *British Journal of Developmental Psychology*, *21*(2), 155–176. https://doi.org/10.1348/026151003765264020

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Amodio, D. M., Harmon-Jones, E., & Devine, P. G. (2003). Individual differences in the activation and control of affective race bias as assessed by startle eyeblink response and self-report. *Journal of Personality and Social Psychology*, *84*(4), 738–753. https://doi.org/10.1037/0022-3514.84.4.738

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

Baker, P. (2018, March 17). Trump and the Truth: A President Tests His Own Credibility. *The New York Times*. https://www.nytimes.com/2018/03/17/us/politics/trump-truth-lies.html

Baron, J., & Jost, J. T. (2019). False Equivalence: Are Liberals and Conservatives in the United States Equally Biased? *Perspectives on Psychological Science*, *14*(2), 292–303. https://doi.org/10.1177/1745691618788876

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bogart, S., & Lees, J. (2023). Meta-perception and misinformation. *Current Opinion in Psychology*, *54*, 101717. https://doi.org/10.1016/j.copsyc.2023.101717

Botvinik-Nezer, R., Jones, M., & Wager, T. D. (2023). A belief systems analysis of fraud beliefs following the 2020 US election. *Nature Human Behaviour*, *7*(7), 1106–1119. https://doi.org/10.1038/s41562-023-01570-4

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2024). Cross-Country Trends in Affective Polarization. *Review of Economics and Statistics*, 1–9. https://doi.org/10.1162/rest_a_01160

Brandt, M. J., Reyna, C., Chambers, J. R., Crawford, J. T., & Wetherell, G. (2014). The Ideological-Conflict Hypothesis: Intolerance Among Both Liberals and Conservatives. *Current Directions in Psychological Science*, *23*(1), 27–34. https://doi.org/10.1177/0963721413510932

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*(3), 389–411. https://doi.org/10.1037/met0000159

Bullock, J., Gerber, A., Hill, S., & Huber, G. (2013). *Partisan Bias in Factual Beliefs about Politics* (w19080; p. w19080). National Bureau of Economic Research. https://doi.org/10.3386/w19080

Carlin, R. E., & Love, G. J. (2018). Political Competition, Partisanship and Interpersonal Trust in Electoral Democracies. *British Journal of Political Science*, *48*(1), 115–139. https://doi.org/10.1017/S0007123415000526

Charness, G., Oprea, R., & Yuksel, S. (2021). How do People Choose Between Biased Information Sources? Evidence from a Laboratory Experiment. *Journal of the*

European Economic Association, *19*(3), 1656–1691. https://doi.org/10.1093/jeea/jvaa051

Clemm Von Hohenberg, B., & Guess, A. M. (2023). When Do Sources Persuade? The Effect of Source Credibility on Opinion Change. *Journal of Experimental Political Science*, *10*(3), 328–342. https://doi.org/10.1017/XPS.2022.2

Cohen, G. L. (2003). Party Over Policy: The Dominating Impact of Group Influence on Political Beliefs. *Journal of Personality and Social Psychology*, *85*(5), 808–822. https://doi.org/10.1037/0022-3514.85.5.808

Cuddy, A. J. C., Fiske, S. T., Kwan, V. S. Y., Glick, P., Demoulin, S., Leyens, J., Bond, M. H., Croizet, J., Ellemers, N., Sleebos, E., Htun, T. T., Kim, H., Maio, G., Perry, J., Petkova, K., Todorov, V., Rodríguez-Bailón, R., Morales, E., Moya, M., … Ziegler, R. (2009). Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology*, *48*(1), 1–33. https://doi.org/10.1348/014466608X314935

Dahlke, R., & Hancock, J. (2022). *The Effect of Online Misinformation Exposure on False Election Beliefs* [Preprint]. Open Science Framework. https://doi.org/10.31219/osf.io/325tn

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*(1), 67–80. https://doi.org/10.1007/s00199-006-0153-z

Derreumaux, Y., Bergh, R., & Hughes, B. L. (2022). Partisan-motivated sampling: Re-examining politically motivated reasoning across the information processing stream. *Journal of Personality and Social Psychology*, *123*(2), 316–336. https://doi.org/10.1037/pspi0000375

Derreumaux, Y., Shamsian, K., & Hughes, B. L. (2023). Computational underpinnings of partisan information processing biases and associations with depth of cognitive reasoning. *Cognition*, *230*, 105304. https://doi.org/10.1016/j.cognition.2022.105304

DeVerna, M. R., Guess, A. M., Berinsky, A. J., Tucker, J. A., & Jost, J. T. (2022). Rumors in Retweet: Ideological Asymmetry in the Failure to Correct Misinformation. *Personality and Social Psychology Bulletin*, 014616722211142. https://doi.org/10.1177/01461672221114222

Di Tella, R., Perez-Truglia, R., Babino, A., & Sigman, M. (2015). Conveniently Upset: Avoiding Altruism by Distorting Beliefs about Others' Altruism. *American Economic Review*, *105*(11), 3416–3442. https://doi.org/10.1257/aer.20141409

Ditto, P. H., Liu, B. S., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., Celniker, J. B., & Zinger, J. F. (2019). At Least Bias Is Bipartisan: A Meta-Analytic Comparison of Partisan Bias in Liberals and Conservatives. *Perspectives on Psychological Science*, *14*(2), 273–291. https://doi.org/10.1177/1745691617746796

Dovidio, J., Hewstone, M., Glick, P., & Esses, V. (2010). *The SAGE Handbook of Prejudice, Stereotyping and Discrimination*. SAGE Publications Ltd. https://doi.org/10.4135/9781446200919

Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., & Ryan, J. B. (2021). How Affective Polarization Shapes Americans' Political Beliefs: A Study of Response to the COVID-19 Pandemic. *Journal of Experimental Political Science*, *8*(3), 223–234. https://doi.org/10.1017/XPS.2020.28

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

Feinberg, M., & Willer, R. (2013). The Moral Roots of Environmental Attitudes. *Psychological Science*, *24*(1), 56–62. https://doi.org/10.1177/0956797612449177

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902. https://doi.org/10.1037/0022-3514.82.6.878

Flynn, D. J., Nyhan, B., & Reifler, J. (2017). The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics. *Political Psychology*, *38*(S1), 127–150. https://doi.org/10.1111/pops.12394

Fridman, A., Gershon, R., & Gneezy, A. (2021). COVID-19 and vaccine hesitancy: A longitudinal study. *PLOS ONE*, *16*(4), e0250123. https://doi.org/10.1371/journal.pone.0250123

Gentzkow, M., Wong, M. B., & Zhang, A. T. (n.d.). *Ideological Bias and Trust in Information Sources*.

Gershman, S. J. (2019). How to never be wrong. *Psychonomic Bulletin & Review*, *26*(1), 13–28. https://doi.org/10.3758/s13423-018-1488-8

Grace, D. M., David, B. J., & Ryan, M. K. (2008). Investigating Preschoolers' Categorical Thinking About Gender Through Imitation, Attention, and the Use of Self-Categories. *Child Development*, *79*(6), 1928–1941. https://doi.org/10.1111/j.1467-8624.2008.01234.x

Green, D. P., Palmquist, B., & Schickler, E. (2002). *Partisan hearts and minds: Political parties and the social identities of voters*. Yale University Press.

Guay, B., & Johnston, C. D. (2022). Ideological Asymmetries and the Determinants of Politically Motivated Reasoning. *American Journal of Political Science*, *66*(2), 285–301. https://doi.org/10.1111/ajps.12624

Hackel, L. M., Mende-Siedlecki, P., & Amodio, D. M. (2020). Reinforcement learning in social interaction: The distinguishing role of trait inference. *Journal of Experimental Social Psychology*, *88*, 103948. https://doi.org/10.1016/j.jesp.2019.103948

Hameleers, M., Brosius, A., & De Vreese, C. H. (2022). Whom to trust? Media exposure patterns of citizens with perceptions of misinformation and disinformation related to the news media. *European Journal of Communication*, *37*(3), 237–268. https://doi.org/10.1177/02673231211072667

Hamman, J. R., Loewenstein, G., & Weber, R. A. (2010). Self-Interest through Delegation: An Additional Rationale for the Principal-Agent Relationship. *American Economic Review*, *100*(4), 1826–1846. https://doi.org/10.1257/aer.100.4.1826

Hansen, G. J., & Kim, H. (2011). Is the Media Biased Against Me? A Meta-Analysis of the Hostile Media Effect Research. *Communication Research Reports*, *28*(2), 169–179. https://doi.org/10.1080/08824096.2011.565280

Hernandez, P., & Minor, D. (2015). Political Identity and Trust. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2635616

Hogg, M. A., Terry, D. J., & White, K. M. (1995). A Tale of Two Theories: A Critical Comparison of Identity Theory with Social Identity Theory. *Social Psychology Quarterly*, *58*(4), 255. https://doi.org/10.2307/2787127

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The Origins and Consequences of Affective Polarization in the United States. *Annual Review of*

Political Science, *22*(1), 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034

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

Kahan, D. M. (2016). The Politically Motivated Reasoning Paradigm, Part 1: What Politically Motivated Reasoning Is and How to Measure It. In R. A. Scott & S. M. Kosslyn (Eds.), *Emerging Trends in the Social and Behavioral Sciences* (1st ed., pp. 1–16). Wiley. https://doi.org/10.1002/9781118900772.etrds0417

Kahn. (2021, May 24). 53% of Republicans view Trump as true U.S. president -Reuters/Ipsos. *Reuters*. https://www.reuters.com/world/us/53-republicans-view-trump-true-us-president-reutersipsos-2021-05-24/

Kim, M. J., Mende-Siedlecki, P., Anzellotti, S., & Young, L. (2021). Theory of Mind Following the Violation of Strong and Weak Prior Beliefs. *Cerebral Cortex*, *31*(2), 884–898. https://doi.org/10.1093/cercor/bhaa263

Kim, M., Park, B., & Young, L. (2020). The Psychology of Motivated versus Rational Impression Updating. *Trends in Cognitive Sciences*, *24*(2), 101–111. https://doi.org/10.1016/j.tics.2019.12.001

Laloggia, J. (n.d.). Republicans and Democrats agree: They can't agree on basic facts. *Pew Research Center*. Retrieved 25 October 2023, from https://www.pewresearch.org/short-reads/2018/08/23/republicans-and-democrats-agree-they-cant-agree-on-basic-facts/

Leeper, T. J., & Slothuus, R. (2014). Political Parties, Motivated Reasoning, and Public Opinion Formation. *Political Psychology*, *35*(S1), 129–156. https://doi.org/10.1111/pops.12164

Leong, Y. C., & Zaki, J. (2018). Unrealistic optimism in advice taking: A computational account. *Journal of Experimental Psychology: General*, *147*(2), 170–189. https://doi.org/10.1037/xge0000382

Levendusky, M. (2013). Partisan Media Exposure and Attitudes Toward the Opposition. *Political Communication*, *30*(4), 565–581. https://doi.org/10.1080/10584609.2012.737435

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

Mende-Siedlecki, P. (2018). Changing our minds: The neural bases of dynamic impression updating. *Current Opinion in Psychology*, *24*, 72–76. https://doi.org/10.1016/j.copsyc.2018.08.007

Mernyk, J. S., Pink, S. L., Druckman, J. N., & Willer, R. (2022). Correcting inaccurate metaperceptions reduces Americans' support for partisan violence. *Proceedings of the National Academy of Sciences*, *119*(16), e2116851119. https://doi.org/10.1073/pnas.2116851119

Michailidou, A., & Trenz, H.-J. (2021). Rethinking journalism standards in the era of post-truth politics: From truth keepers to truth mediators. *Media, Culture & Society*, *43*(7), 1340–1349. https://doi.org/10.1177/01634437211040669

Mothes, C., & Ohme, J. (2019). Partisan Selective Exposure in Times of Political and Technological Upheaval: A Social Media Field Experiment. *Media and Communication*, *7*(3), 42–53. https://doi.org/10.17645/mac.v7i3.2183

Park, B., Fareri, D., Delgado, M., & Young, L. (2021). The role of right temporoparietal junction in processing social prediction error across relationship contexts. *Social Cognitive and Affective Neuroscience*, *16*(8), 772–781. https://doi.org/10.1093/scan/nsaa072

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

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

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

Peterson, E., & Iyengar, S. (2021). Partisan Gaps in Political Information and Information-Seeking Behavior: Motivated Reasoning or Cheerleading? *American Journal of Political Science*, *65*(1), 133–147. https://doi.org/10.1111/ajps.12535

Pronin, E., Lin, D. Y., & Ross, L. (2002). The Bias Blind Spot: Perceptions of Bias in Self Versus Others. *Personality and Social Psychology Bulletin*, *28*(3), 369–381. https://doi.org/10.1177/0146167202286008

Rathje, S., Roozenbeek, J., Van Bavel, J. J., & Van Der Linden, S. (2023). Accuracy and social motivations shape judgements of (mis)information. *Nature Human Behaviour*, *7*(6), 892–903. https://doi.org/10.1038/s41562-023-01540-w

Roozenbeek, J., Culloty, E., & Suiter, J. (2023). Countering Misinformation: Evidence, Knowledge Gaps, and Implications of Current Interventions. *European Psychologist*, *28*(3), 189–205. https://doi.org/10.1027/1016-9040/a000492

Roozenbeek, J., Freeman, A. L. J., & Van Der Linden, S. (2021). How Accurate Are Accuracy-Nudge Interventions? A Preregistered Direct Replication of Pennycook et al. (2020). *Psychological Science*, *32*(7), 1169–1178. https://doi.org/10.1177/09567976211024535

Rutjens, B. T., Sutton, R. M., & Van Der Lee, R. (2018). Not All Skepticism Is Equal: Exploring the Ideological Antecedents of Science Acceptance and Rejection. *Personality and Social Psychology Bulletin*, *44*(3), 384–405. https://doi.org/10.1177/0146167217741314

Schulz, L., Schulz, E., Bhui, R., & Dayan, P. (2023). *Mechanisms of Mistrust: A Bayesian Account of Misinformation Learning* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/8egxh

Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences*, *108*(19), 7710–7715. https://doi.org/10.1073/pnas.1014345108

Swire, B., Berinsky, A. J., Lewandowsky, S., & Ecker, U. K. H. (2017). Processing political misinformation: Comprehending the Trump phenomenon. *Royal Society Open Science*, *4*(3), 160802. https://doi.org/10.1098/rsos.160802

Taber, C. S., & Lodge, M. (2006). Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science*, *50*(3), 755–769. https://doi.org/10.1111/j.1540-5907.2006.00214.x

Tajfel, H., & Turner, J. C. (2004). The Social Identity Theory of Intergroup Behavior. In J. T. Jost & J. Sidanius (Eds.), *Political Psychology* (0 ed., pp. 276–293). Psychology Press. https://doi.org/10.4324/9780203505984-16

Thaler, M. (2020). The fake news effect: Experimentally identifying motivated reasoning using trust in news. *arXiv Preprint arXiv:2012.01663*.

The Washington Post. (2021). *Tracking all of President Trump's false or misleading claims—Washington Post*. https://www.washingtonpost.com/graphics/politics/trump-claims-database/

Traast, I. J., Amodio, D., Schultner, D., & Doosje, B. (2023). *Race effects on impression formation in social interaction: An instrumental learning account*. https://osf.io/rnjgh/

Turner, J. C., Oakes, P. J., Haslam, S. A., & McGarty, C. (1994). Self and Collective: Cognition and Social Context. *Personality and Social Psychology Bulletin*, *20*(5), 454–463. https://doi.org/10.1177/0146167294205002

Vallone, R. P., Ross, L., & Lepper, M. R. (1985). The hostile media phenomenon: Biased perception and perceptions of media bias in coverage of the Beirut massacre. *Journal of Personality and Social Psychology*, *49*(3), 577–585. https://doi.org/10.1037/0022-3514.49.3.577

Van Bavel, J. J., & Pereira, A. (2018). The Partisan Brain: An Identity-Based Model of Political Belief. *Trends in Cognitive Sciences*, *22*(3), 213–224. https://doi.org/10.1016/j.tics.2018.01.004

Van Der Linden, S. (2022). Misinformation: Susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, *28*(3), 460–467. https://doi.org/10.1038/s41591-022-01713-6

Van Der Linden, S., Panagopoulos, C., Azevedo, F., & Jost, J. T. (2021). The Paranoid Style in American Politics Revisited: An Ideological Asymmetry in Conspiratorial Thinking. *Political Psychology*, *42*(1), 23–51. https://doi.org/10.1111/pops.12681

Vermue, M., Meleady, R., & Seger, C. R. (2019). Member-to-member generalisation in trust behaviour: How do prior experiences inform prosocial behaviour towards novel ingroup and outgroup members? *Current Psychology*, *38*(4), 1003–1020. https://doi.org/10.1007/s12144-019-00289-8

Wagner, M. (2021). Affective polarization in multiparty systems. *Electoral Studies*, *69*, 102199. https://doi.org/10.1016/j.electstud.2020.102199

Wolsko, C., Ariceaga, H., & Seiden, J. (2016). Red, white, and blue enough to be green: Effects of moral framing on climate change attitudes and conservation behaviors. *Journal of Experimental Social Psychology*, *65*, 7–19. https://doi.org/10.1016/j.jesp.2016.02.005

Xiao, Y. J., Coppin, G., & Van Bavel, J. J. (2016). Perceiving the World Through Group-Colored Glasses: A Perceptual Model of Intergroup Relations. *Psychological Inquiry*, *27*(4), 255–274. https://doi.org/10.1080/1047840X.2016.1199221

Zimmermann, F. (2020). The Dynamics of Motivated Beliefs. *American Economic Review*, *110*(2), 337–363. https://doi.org/10.1257/aer.20180728