# Individual differences in vulnerability to misinformation when forming impressions of political candidates

Michael S. Cohen[1,2 *], Victoria Halewicz[1,3], Ece Yildirim[1], Joseph W. Kable[1 *]

[1]Department of Psychology, University of Pennsylvania, Philadelphia, PA

[2]Department of Psychology, University of Chicago, Chicago, IL

[3]Department of Psychology, Brown University, Providence, RI

**Keywords:** continued influence effect, misinformation, analytic thinking, digital literacy, political partisanship

*Address correspondence to:
Michael S. Cohen
Department of Psychology
University of Chicago
5848 S. University Ave.
Chicago, IL 60637
E-mail: mscohen@uchicago.edu

Joseph W. Kable
Department of Psychology
University of Pennsylvania
3700 Hamilton Walk
Philadelphia, PA 19104
E-mail: kable@psych.upenn.edu

**Abstract**

Previous work has shown that false information continues to affect decision making even after being corrected, a phenomenon known as "continued influence effects" (CIEs). Here we demonstrate that vulnerability to CIEs varies systematically between individuals as a function of demographic and psychometric variables. We developed a set of mock social media accusations, refutations, and control stimuli targeting fictional political candidates. We observe robust within-participant CIEs: candidates targeted by corrected accusations evoke lower feeling thermometer ratings than candidates not targeted by accusations. Individuals who rely more on intuitive feelings show larger CIEs, whereas those who score higher on digital literacy show reduced CIEs. These results suggest that analytic thinking plays a role in countering the continued influence of corrected misinformation. Interestingly, older adults appear less vulnerable to CIEs than their younger counterparts, a counterpoint to prior findings that older adults share more false content on social media. We find no effect of political orientation on CIEs despite its influence on explicit identification of misinformation. Finally, after a two-day delay, accusation stimuli are remembered better than refutations, suggesting that accusations stimulate higher-priority processing than refutations, potentially due to stronger emotional arousal. Our results suggest that analytic thinking and digital literacy could be protective when people must judge political candidates targeted by refuted false information.

**Significance Statement**

False information, even after being corrected, can still influence subjective impressions and decisions about its targets. We address this issue using a novel approach: presenting mock social media posts regarding a large set of mock political candidates. Some candidates are targeted by an allegation, and some of the allegations are factually refuted. We find that individuals who rely more on intuitive thinking or who have lower digital literacy are more vulnerable to the influence of corrected false information. However, political partisanship does not impact vulnerability. Since real-world harms from misinformation occur primarily via distortion of decision-making, our approach elucidates who is most vulnerable to making poor decisions based on false information.

## Introduction

Content evoking strong moral-emotional reactions tends to receive greater social media engagement (1). False content tends to evoke more negative emotions than true content (2), resulting in rapid distribution in a social media ecosystem. In an ideal world, factual refutations would counter the influence of false content. In actuality, even when people become aware of the falsehood, inaccurate information can still influence their later judgments, a phenomenon known as "continued influence effects" (CIEs), e.g., (3, 4, 5). Past work has documented the existence of CIEs and has examined underlying cognitive mechanisms, but has not yet considered what contributes to individual-level susceptibility. We introduce a novel approach that measures CIEs more precisely, enabling an examination of demographic and psychometric variables that predict the magnitude of CIEs between individuals.

CIEs were originally studied in the context of causal reasoning (3), for example, about the causes of a fire. More recent work has shown that CIEs occur in the political realm, where refuted misconduct allegations still result in negative evaluations (4). Other work has explored the effectiveness of corrections for CIEs and found that providing full factual refutations is more successful than merely asserting that an accusation is false (5). We build on this prior work to examine which individual-level factors make people vulnerable to persistent effects of refuted misinformation when making sociopolitical decisions.

### Factors predicting explicit belief in and sharing of misinformation

Though little work has examined predictors of vulnerability to CIEs, previous studies have examined factors that predict which individuals are most likely to have difficulty identifying misinformation, to believe in conspiracy theories, and to share misinformation. We expect that some of these factors may similarly predict vulnerability to CIEs.

Prior work has suggested that a person's ability to think analytically rather than intuitively plays an important role in reducing vulnerability to misinformation. Specifically, higher scores on the Cognitive Reflection Test (CRT), a measure of the ability to use analytical reasoning to overcome intuitive but incorrect responses, predicts greater accuracy discernment ─ the ability to identify false information (6). Another factor that predicts improved accuracy discernment is factual knowledge about digital and legacy media such as content moderation on online platforms (7, 8); this knowledge may enhance the ability to think analytically about online content. A self-report measure of epistemic belief also correlates with the tendency to believe specific conspiracy theories (9). Specifically, of the three subscales on this measure, having faith in one's intuitive feelings and believing that truth is political (i.e., defined by those in power) are associated with greater levels of belief in common conspiracy theories, whereas requiring evidence as a basis for beliefs is associated with reduced endorsement of conspiracy theories.

In contrast, other work has suggested that ideological biases play an important and potentially greater role in vulnerability to misinformation (10). Specifically, Republicans and those scoring lower on actively open-minded thinking (AOT) show lower accuracy discernment, suggesting a tendency for ideological bias known as "myside bias". Partisanship and AOT are stronger predictors of accuracy discernment than measures of analytic thinking (11), supporting an "integrative account" of misinformation vulnerability. Another measure conceptually related to ideological bias is affective polarization, a strong preference for individuals within one's party over opposing political partisans. Higher levels of affective polarization have been associated with a greater likelihood of sharing false content on Twitter, particularly among Republicans (12).

Finally, past work has been inconclusive as to whether older adults are more or less vulnerable to misinformation. On one hand, older adults consume and share higher levels of false content on real social media platforms (13, 14). However, other work shows that older adults are better at discerning true from false content in laboratory settings (15, 16). Additional evidence will help clarify these seemingly conflicting findings.

**Goals of the present study**

Past studies examining individual differences in vulnerability to misinformation primarily measure explicit decisions about the false content itself. However, it is not clear whether individuals who can discern false information can also avoid the impact of corrected misinformation on subsequent decisions. We aim to determine whether the same factors that predict the likelihood of explicitly identifying false information also predict the likelihood of ignoring debunked information when making decisions.

Additionally, we aim to achieve a preliminary understanding of differences in processing accusations versus factual refutations. Specifically, we examined whether recognition memory varies between accusations and refutations. Prior work suggests that negative information is motivationally salient in a variety of evaluative contexts, including in social impression formation (17), and that negative emotion enhances memory for associated stimuli (18). We hypothesized that accusations benefit from prioritization relative to neutral stimuli, while refutations benefit less or not at all. Separate from any direct effect of memory modulation on decision making, prioritized processing of accusation stimuli is a possible mechanism by which refuted accusations could still bias decisions (cf., 19).

## Results

We ran two large online behavioral experiments, aiming to recruit 500 participants per experiment. The first study was exploratory, while the second was a preregistered replication in which candidate judgements were also measured two days later. In both experiments, participants first read introductory bios for all candidates (Figure 1A), and then saw two mock social media posts for each candidate (Figure 1B). Each set of social media posts was presented in one of three formats, which were varied within-participants: *corrected accusation* (Post 1: Accusation; Post 2: Refutation), *uncorrected accusation* (Post 1: Accusation; Post 2: Neutral), or *no accusation* (Post 1: Neutral; Post 2: Neutral). Participants saw an equal number of candidates in each of the three formats, and which candidates appeared in which format was counterbalanced across participants. Immediately after reading the posts, participants rated the candidate on a feeling thermometer (Figure 1C). Three distinct comparisons can be made between conditions: 1) Continued influence effects (CIEs), subtracting mean ratings for candidates with *no accusation* from mean ratings for candidates with *corrected accusations*; 2) Accusation effects, subtracting mean ratings for candidates with *no accusation* from mean ratings for candidates with *uncorrected accusations*, and 3) Correction effects, subtracting mean ratings for candidates with *uncorrected accusations* from mean ratings for candidates with *corrected accusations*. CIEs are the primary focus of the analyses presented below; analogous data for accusation effects and correction effects on immediate ratings are reported in Supplemental Tables S1-S2. Note that more negative scores on CIEs indicate a larger decline in ratings for candidates exposed to accusations.

After reading posts about each candidate and providing a rating, participants then completed a series of self-report questionnaires (see Methods), as well as the MIST headline

accuracy judgment measure (20). These were followed by delayed ratings about each candidate, following the same procedure as the immediate ratings, as well as a delayed choice task (Figure 1D), in which participants completed a series of binary choices about which of two candidates they would vote for in an election. In the choice task, we computed condition effects as the proportion of choices favoring a given candidate type over a comparison candidate type. In other words, we computed CIEs as the proportion of trials where a corrected accusation candidate was chosen over a no accusation candidate when candidates of those two types were being compared. Finally, memory was assessed via a recognition memory test (Figure 1E). In Experiment 1, all of these measures were collected in a single testing session, while in Experiment 2, memory and a second round of delayed ratings and choices were collected two days later.

**Behavioral main effects**

**Immediate and delayed CIEs.** As expected based on pilot testing (see SI Results), CIEs were present in the aggregate in Experiment 1 on immediate ratings, $t(457) = -11.58$, $p < .001$, $d = -.54$ (Figure 2A). CIEs remained present on ratings made after a short (~20-30 minute) delay, $t(457) = -5.30$, $p < .001$, $d = -.25$ (Figure 2B). CIEs were also present on the choice task at a short delay in Experiment 1, $t(457) = -5.09$, $p < .001$, $d = -.24$ (Figure 2C). In preregistered analyses for Experiment 2, CIEs were present on immediate ratings, $t(500) = -11.47$, $p < .001$, $d = -.51$ (Figure 2D), on ratings made after a short delay, $t(500) = -5.53$, $p < .001$, $d = -.25$ (Figure 2E), and on choices made after a short delay, $t(500) = -4.58$, $p < .001$, $d = -.20$ (Figure 2F). In Experiment 2, we were also able to examine CIEs after a two-day delay; these effects were present in pre-registered analyses for both the delayed rating task, $t(423) = -5.26$, $p < .001$, $d = -.26$, and the delayed choice task, $t(423) = -4.69$, $p < .001$, $d = -.23$ (see Figure S1). Contrary to our preregistered prediction, CIEs in Experiment 2 did not differ between short-delay and long-

delay measurements for the rating task, $t(423) < 1$, $d = .04$, or for the choice task, $|t|(423) < 1$, $d = -.02$ (this issue is discussed further in SI Results).

**CIEs and memory.** We additionally find that CIEs do not appear to be a direct consequence of trials in which the accusation was remembered while the refutation was not. That is, CIEs were still present when limiting the analysis to trials in which all four stimulus types (cf. Fig. 1B) were correctly categorized as previously seen or not on the later recognition test, though many participants needed to be excluded from these analyses due to having no such trials in a given condition. CIEs were present on immediate ratings when conditionalizing on memory in Experiment 1, $t(332) = -5.81$, $p < .001$, $d = -.32$, and in Experiment 2, $t(222) = -6.10$, $p < .001$, $d = -.41$. CIEs conditionalized on memory were also present on short-delay ratings in Experiment 1, $t(332) = -5.85$, $p < .001$, $d = -.32$, on short-delay ratings in Experiment 2, $t(222) = -4.50$, $p < .001$, $d = -.30$, and on long-delay ratings in Experiment 2, $t(222) = -4.92$, $p < .001$, $d = -.33$.

Data from the choice task were less clear. Here, we limited the analysis to candidate pairings in which all four stimulus types were remembered correctly for both candidates, and to individuals for whom at least three choice trials relevant to CIEs met this criterion. This analysis shows a significant CIE at a short delay in Experiment 1, $t(196) = -2.85$, $p = .005$, $d = -.20$, but not in Experiment 2, whether at a short delay, $t(49) = -1.47$, $p = .147$, $d = -.21$, or at a long delay, $|t|(49) < 1$, $d = -.12$. Still, all effects are in the same direction as in the analysis not conditionalized on memory, with corrected accusation candidates less likely to be chosen than no accusation candidates, and with effect sizes that are comparable to the main analysis. To more systematically test whether the observed null effects may be due to a failure to remember the refutation, we compared whether the proportion of corrected accusation candidates chosen

differs within the same individuals when the analyses are or are not conditionalized on memory. No evidence for such a difference was found, whether in Experiment 1, $|t|(196) < 1$, $d = -.06$, in Experiment 2 at a short delay, $|t|(49) < 1$, $d = -.05$, or in Experiment 2 at a long delay, $|t|(49) < 1$, $d = -.05$. Thus, the null effects for choices in Experiment 2 are more likely due to a lack of power when conditionalizing on successful memory after a two-day delay; we find no direct evidence supporting a role for memory failures in producing CIEs.

**Stepwise regressions**

We used a stepwise regression (see "Data analysis" in Methods) to examine which factors predicted CIEs on the immediate feeling thermometer measure. We focus on immediate ratings because CIEs are largest for this measure, creating a more reliable signal in which to examine individual differences. Results of analogous analyses on ratings and choices made after a delay, and of models directly comparing effects of predictor variables on immediate and delayed outcome measures, are presented in Supplemental Tables S3-S4 (CIEs) and Supplemental Tables S5-S8 (accusation and correction effects).

Four variables were significant predictors of immediate CIEs in Experiment 1 (Table 1A): Higher faith in intuition and greater affective polarization predicted stronger CIEs, while older age and higher digital literacy predicted weaker CIEs. A comparable analysis was preregistered for Experiment 2, and results from Experiment 1 partially replicated (Table 1B): faith in intuition again significantly predicted stronger CIEs, older age significantly predicted weaker CIEs, and digital literacy predicted marginally weaker CIEs. In a follow-up exploratory analysis for Experiment 2, we removed the affective polarization variable, and included 56 political independents who we had originally excluded due to the inability to compute affective polarization for true independents. Here, digital literacy ($\beta = 0.118$, $t = 2.49$, $p = .013$) was a

significant predictor, alongside age (β = 0.131, $t$ = 2.84, $p$ = .005) and epistemic faith in intuition (β = -0.134, $t$ = -2.87, $p$ = .004), with gender also included in the model as a nonsignificant predictor (β = -0.071, $t$ = -1.56, $p$ = .119).

Contrary to our predictions, affective polarization was not a significant predictor in Experiment 2. Hence, our preregistered analysis of how the effects of affective polarization differ based on the relationship between participant and candidate ideology is reported in Supplemental Table S9. While the reason for the difference between experiments is not clear, an additional exploratory analysis yields some insight. Specifically, when affective polarization and its interaction with political party are added to the models emerging from the stepwise regression in Table 1, the interaction effect is significant in Experiment 2 (β = -0.108, $t$ = -2.19, $p$ = .029) but not in Experiment 1 (β = 0.027, $t$ = 0.53, $p$ = .60). In other words, in Experiment 2, the relationship between affective polarization and greater CIEs was stronger among Republicans, while in Experiment 1, this relationship was consistently found across all participants.

**CIEs vs. Headline accuracy discernment**

We next examined whether the regression coefficients for any variables that predict CIEs reliably differ from those that predict accuracy discernment. To address this question, we used the SUR method (21), testing all significant or marginal variables ($p < .10$) from either of the analyses being compared. In Experiment 1, two predictors showed significant differences between outcome measures after correction for multiple comparisons: political party and affective polarization (Table 2A). Specifically, Republicans scored worse on headline accuracy discernment, but there was no effect of party on CIEs. Greater affective polarization predicted significantly larger CIEs, but the trend for accuracy discernment was in the opposite direction. A preregistered analysis in Experiment 2 (Table 2B) showed that the differential effect of political

party replicates. The differential effect of affective polarization, however, did not replicate. Affective polarization did not meet the criteria for inclusion in an SUR model in Experiment 2, but even if it were added to the model described in Table 2B, there was no evidence for a difference between outcome measures in Experiment 2, as uncorrected $p = .73$. Finally, of the additional exploratory measures added in Experiment 2, only Actively Open-minded Thinking (AOT) was a significant predictor of either MIST or CIE scores. When adding AOT to the models shown in Table 2B, it was a significant predictor of MIST scores ($\beta = 0.187$, $t = 4.09$), but not of CIE ($\beta = -0.014$, $t = -0.26$); this difference between measures is significant, $\chi^2 = 8.04$, $p = .0046$, without multiple comparison correction.

**Memory**

Finally, we examined whether explicit memory differs for accusations versus refutations. Although CIEs are still present when participants remember all information, memory differences could nonetheless provide insight into how accusations and refutations are processed differently. We specifically examined whether the categorical benefit to memory for stimuli that should impact social impressions (i.e., accusations and refutations), relative to matched neutral stimuli lacking such impact, would differ for accusations vs. refutations. To do so, we ran a 2 (Impactfulness: Accusation/Refutation vs. Control) x 2 (Post: Post 1 vs. Post 2) repeated-measures ANOVA. In Experiment 1, this analysis showed a main effect of impactfulness, $F(1, 437) = 27.32$, $p < .001$, $\eta_p^2 = .06$, with impactful stimuli generally being remembered better than neutral stimuli, and a marginal interaction between impactfulness and post, $F(1, 437) = 3.36$, $p = .067$, $\eta_p^2 = .01$, demonstrating a slight tendency for this effect to be larger for accusations than for refutations (Figure 3A).

We examined this effect again in Experiment 2, in which memory was tested 2 days after encoding rather than ~30-40 minutes after encoding (Figure 3B). Because emotional stimuli tend to affect memory consolidation more than immediate memory (22), we preregistered a prediction of a reliable interaction effect in Experiment 2. Indeed, in Experiment 2, we found a main effect of impactfulness, $F(1, 404) = 23.07$, $p < .001$, $\eta_p^2 = .05$, a main effect of post, $F(1, 404) = 31.18$, $p < .001$, $\eta_p^2 = .07$, and critically, an interaction between these variables, $F(1, 404) = 26.34$, $p < .001$, $\eta_p^2 = .06$. The interaction indicates a strong memory benefit for accusation stimuli vs. neutral post 1 stimuli, $t(404) = 7.14$, $p < .001$, $d = .36$, but no advantage for refutation stimuli vs. neutral post 2 stimuli, $|t|(404) < 1$, $d = -.03$. Finally, we ran an additional 2 x 2 x 2 (Impactfulness x Post x Experiment) mixed ANOVA to test whether the interaction between impactfulness and post differed as a function of the retention interval. This analysis showed a 3-way interaction, $F(1, 841) = 7.78$, $p = .005$, $\eta_p^2 = .01$, indicating that the memory benefit for accusation stimuli was greater with a longer retention interval, as well as a main effect of experiment, $F(1, 841) = 138.08$, $p < .001$, $\eta_p^2 = .14$, reflecting poorer memory at the longer retention interval, an interaction between post and experiment, $F(1, 841) = 11.22$, $p < .001$, $\eta_p^2 = .01$, and other lower-order effects repeating those reported above.

## Discussion

In this study, we validate a novel approach to studying the persistent negative impact of misconduct allegations on impressions of political candidates, even after those allegations have been factually refuted. Specifically, we find that candidates targeted by refuted accusations are rated more poorly than those who were never accused, replicating prior work on continued influence effects (CIEs). This effect persists regardless of whether ratings are made immediately, after a short delay, or after a two-day delay. We additionally find reliable individual differences

in CIEs. CIE magnitude is related to reliance on intuition, digital literacy, and age, while it is notably unrelated to political affiliation. Finally, while we do not find any evidence to suggest that CIEs are a direct result of selective forgetting of refutation stimuli, we do find that participants remember accusations better than refutations, particularly after a two-day delay. We speculate that memory modulation results from affective processes that simultaneously alter both memory encoding/consolidation processes and decision making about targeted candidates.

The finding of robust CIEs in our paradigm largely replicates a prior study in which accusations target mock political candidates (4). It contrasts with another recent study in which CIEs did not occur in social impression formation towards hypothetical fellow students (23). Others (24) have suggested based on the juxtaposition between (4) and (23) that people may be less willing to update negative views about disliked politicians than about hypothetical people similar to those they encounter in everyday life. This synthesis was influenced by the finding in (4) that CIEs may be eliminated when decisions are both preceded by explicit deliberation and are about a same-party political candidate. We conclude that in a non-partisan political context, CIEs towards politicians are robust. We also examined whether the relationship between a candidate's perceived ideology and the participant's ideology related to CIEs (see SI Results) and found that ideological alignment between the candidate and participant is not protective, and if anything was associated with increased CIEs. Although results may differ when partisan identity is more salient (our stimuli notably do not contain explicit partisan cues beyond vague policy positions), these findings suggest that CIEs apply more broadly in sociopolitical impression formation than was suggested by (24). Our study is also unique in presenting many targets rather than only a single target, which may reduce cognitive resources available for

impression updating. In any case, we find strong evidence in the present paradigm for robust aggregate CIEs in impression formation toward political candidates.

We find as well that the degree to which retracted information continues to influence social impressions varies between individuals, with people who rely more on intuition and less on analytic thinking showing greater vulnerability to CIEs. Specifically, those who self-report increased reliance on intuition are more likely to be influenced by debunked information. In contrast, individuals with higher digital literacy—an acquired ability to understand how social media platforms work on a mechanistic level—show reduced CIEs. This observation extends prior findings that increased reliance on emotion and intuition, measured by self-reported emotional state or manipulated by encouraging the use of emotions in making decisions, leads to misjudging false information as accurate (25). Thus, intuition and digital literacy show effects on CIEs that are comparable to those observed on accuracy discernment both in our study and in previous work (e.g., 7, 8, 9, 25).

We also find that ideological bias (i.e., partisan orientation) is unrelated to the magnitude of CIEs. At the same time, we replicate prior findings that partisan orientation is associated with poorer headline accuracy discernment, as is actively open-minded thinking (AOT), as suggested by the "integrative account" of misinformation vulnerability (10, 11). Thus, we conclude that although Republicans and those with low-AOT worldviews are more inclined to believe the specific false claims tested in the MIST, they are not less willing than Democrats to update impressions of previously unknown candidates when an accusation is debunked. Our results suggest that these individuals are not inherently more vulnerable to making decisions based on debunked false information. However, American conservatives in recent years consuming content from identity-congruent media and elites are exposed to more false information than

similarly situated liberals (e.g., 13, 26, 27). Participants would have never heard opinions from media or elites about the specific candidates presented in our study. Thus, our data imply that cues from elites and the media about specific candidates or issues, rather than features inherent to one's worldview (cf., 28), may play a critical role in reducing accuracy discernment as measured by the MIST, while not affecting CIEs in our novel paradigm.

We additionally observed that older adults consistently show reduced CIEs. This result aligns with research on the age-related positivity bias (29), including recent work from our lab showing that older adults are more forgiving of selfish behavior than young adults (30). Older adults may similarly be less inclined than young adults to treat a mere accusation of immoral behavior by a hypothetical politician as reflecting poor moral character, particularly once evidence has surfaced to refute the accusation. Our result is also consistent with prior work demonstrating better headline accuracy discernment in older adults (15, 16). Still, in light of these findings, it remains puzzling why older adults share more false information on social media compared to their younger counterparts (13, 14).

Finally, we found evidence that accusations receive preferential processing relative to refutations. People had better memory for accusations than refutations, and this effect was reliably stronger at a two-day delay in Experiment 2, relative to the shorter retention interval in Experiment 1. This result is consistent with enhanced consolidation driven by increased arousal in response to accusations (22). As discussed further in SI Results, we also found tentative evidence, based on the relationship between memory and CIEs in Experiment 2, more directly suggestive of affective factors playing a potential role in CIEs. Still, further work will be needed to more definitively establish how the prioritization of accusation stimuli affects decisions made about targeted candidates.

We expect this work to inspire various lines of future research. Our work is a proof of concept that the degree to which refuted false information continues to affect decisions can be assessed at the individual level, which can then be used to measure who is most vulnerable to false information. By measuring decisions about targets of misinformation, our approach is uniquely suited to provide insights about who is most vulnerable to having consequential decisions distorted by misinformation even after a factual correction. Our approach is also unique in measuring misinformation's influence on decisions over time. This approach will furthermore enable future work aimed at clarifying the mechanisms by which false information influences subjective judgments. It could be adapted to explore misinformation about other types of targets, including false claims of harm from COVID-19 vaccines or false claims of benefits from unproven medical treatments. Ultimately, we see this work as a critical first step towards understanding how misinformation affects decisions, and towards ensuring that interventions mitigate real harms that affect behavior.

## Methods

### Stimuli

Our stimuli were initially developed as a set of 36 political candidates. This was reduced to 27 candidates for the present set of studies to keep the duration of the experiment session under 60 minutes. The full stimulus set is available in both forms (27 and 36 candidates) in our OSF repository (https://osf.io/gjpr9/). Each candidate had a story inspired by a real politician targeted by a false accusation that was later debunked by media sources including FactCheck.org. The stimuli included stories from each of nine scandal categories: bribery, electoral fraud/interference, embezzlement/self-dealing, racism, abuse of power or discretion, sexual harassment, foreign influence, financial fraud, and pedophilia/bestiality. Each political

candidate had a novel face and name. The stimulus set included 18 men and 9 women, 18 white candidates, and 9 non-white candidates split evenly between Black, Hispanic, and Asian race/ethnicity. Candidates were assigned to one of three political offices, with nine candidates running for each office: U.S. Senate, state governor, or U.S. House (sometimes referred to as "running for Congress"). As much as possible, mock candidates resembled the real candidate who inspired the story on race, gender, age, and political office, though some deviation was necessary because white men were overrepresented in the real-life sample relative to our intended distribution.

For each novel candidate, we constructed a set of core stimuli ─ introductory bios along with accusation, refutation, and neutral social media posts. The bios were typically inspired by the original candidate, while avoiding potentially identifiable information. Above each bio was a banner reminiscent of a political bumper sticker with the candidate's face, name, and political office. Each accusation, refutation, or neutral social media post appeared below a candidate banner, and had a blurred stock photo as a thumbnail, with a few lines describing the story and a mock link featuring a thumbnail image and headline that would preview a hypothetical linked article on Facebook.

Each participant saw two social media posts about a given political candidate. The first post contained either an accusation or neutral content while the second post contained either a refutation or neutral content. To minimize possible confounds, the first post had some consistent features whether it was an accusation or a neutral control post: namely, the same image representing the hypothetical linked article (but different text captions), the same blurred thumbnail of the author, and a similar topic area. The second post was also constructed with these features matched regardless of whether it was a refutation or a neutral control post. In the

second post, however, the author's thumbnail and account name were from a blurred, reputable news source (e.g., The New York Times) with an unblurred blue checkmark. The refutation posts provided a clear causal account for the accusation by explaining what information was misinterpreted or who was responsible for the erroneous or deliberately fabricated narrative. See SI Methods for additional details about stimulus design and pilot testing.

**Experiment 1 – Exploratory behavioral study**

**Participants**

We planned to recruit 500 participants using the CloudResearch MTurk Toolkit. Recruitment was stratified by age, with equal samples targeted from five age brackets (18-29, 30-39, 40-49, 50-59, 60+). Participants were paid $8.00 for completing the study. Participants were excluded for poor performance if they scored at or below chance on measures with a clear correct answer (specifically, 50% or less on the MIST-20 or under 25% on the digital literacy measure) or if they failed simple attention checks included at the beginning of the protocol. A total of 499 participants completed the study and 41 participants were excluded, leaving a sample of 458 participants.

**Procedure**

Participants were first presented with introductory bios (Figure 1A) in random order for all 27 candidates. Immediately after each bio, they were asked to rate how much they liked each candidate on a 0-100 feeling thermometer scale (initial ratings). Participants then saw two mock social media posts about each candidate (Figure 1B). Candidates were evenly divided between the three possible conditions (*corrected accusation*, *uncorrected accusation*, or *no accusation*), with assignment of candidate to specific conditions counterbalanced across participants. After

each pair of posts, participants again rated each candidate on a 0-100 feeling thermometer scale (immediate post-story ratings; Figure 1C).

After viewing posts about all 27 candidates, participants completed a series of established questionnaire measures. These were the Cognitive reflection test (CRT-2) (30); epistemic beliefs measure (9); two affective polarization measures – a dictator game, following (32), and a partisan feeling thermometer, following (33); a belief superiority measure (34); and the MIST-20 headline accuracy discernment scale (20). We also included a novel digital literacy measure, building on (7, 8); the digital literacy measure included 6 multiple choice questions with 4 response options each, to examine factual knowledge of user experience and content moderation on social media platforms. More specifically, the digital literacy measure includes three questions about specific platforms (Facebook, Twitter, and TikTok), and three questions about specific concepts (phishing, blocking, and tagging). The percentage correct across all six items was used as a measure of digital literacy. See additional details in SI Results and Appendix.

Following these questionnaires, participants made two judgments about each candidate. First was a choice task, in which participants indicated which of two candidates they would prefer for a given office, with all possible pairings of candidates running for the same political office presented (Figure 1D). Candidates were assigned to one of three offices (Governor, U.S. Senate, or U.S. House), yielding 36 choices per office, and a total of 108 choice trials. During the choice task, a pair of candidate banners were shown together on-screen for each trial. After the choice task, participants completed a delayed feeling thermometer rating for each candidate, again prompted by the candidate banner. This was followed by a recognition memory test (Figure 1E) in which all four stimuli (accusation, accusation control, refutation, refutation control) were shown for each of the 27 candidates, and participants responded using a 5-point

scale ranging from “Definitely new” to “Definitely old”. Finally, participants answered demographic questions about their age, race/ethnicity, party affiliation and political ideology (both following 7-point ANES survey format), education, income, household size, and ZIP code.

**Data analysis**

On feeling thermometer rating measures, we computed a CIE score for each individual by subtracting the individual’s mean score for *no accusation* candidates from the mean score for *corrected accusation* candidates. Note that this means that negative scores indicate the presence of CIEs, with larger CIEs yielding more negative scores. This approach was chosen so that negative values indicate greater influence of false information for both CIEs and headline accuracy discernment. For choices, CIEs were computed as the proportion of trials in which the options were a *corrected accusation* candidate and a *no accusation* candidate in which the *corrected accusation* candidate was chosen. Here again, lower scores indicate greater CIEs, relative to chance performance of 50%. For exploratory analyses examining individual differences in CIEs on delayed feeling thermometer ratings and delayed choices, the Z-score across the sample was computed for each individual on each of the two measures, and the two Z-scores were averaged to yield a composite measure of delayed ratings/choices.

We also computed an affective polarization measure for those who expressed a preference between the Republican and Democratic parties; those with a weak preference were included, but those who expressed no partisan preference were not. We computed scores on this measure by taking the differences in scores for in-party vs. out-party targets on dictator game offers (how much was shared out of $10) and feeling thermometer ratings. For each measure, the Z-score of the in-party vs. out-party difference was computed across the sample, and the final measure of affective polarization was the participant’s average Z-score across the two measures.

For demographic variables of education and income, we converted categorical responses to ordinal numbers for inclusion in regressions, as described in our preregistration. Education was converted to years of education such that “Doctoral degree” = 20, “Master’s degree” = 18, “Bachelor’s degree” = 16, “Associate’s Degree” = 14, “Some college” = 13, “High school diploma” = 12, “Have not finished high school” = 11. A small number of respondents chose “Other”, and these responses were coded on an ad hoc basis, with vocational/technical school graduates and those currently in college coded as 13, and a respondent indicating a professional degree coded as 19. Income was converted to a category, consistent with our preregistered analysis plan, with “Under $20,000” = 1, “$20,000-$40,000” = 2, “$40,000 - $75,000” = 3, “$75,000-$100,000” = 4, “$100,000-$500,000” = 5, and “Over $500,000” = 6.

Our primary regression analysis was a stepwise regression, run using the stepAIC algorithm in the R MASS package. The following variables were included in the initial model: CRT % correct, Digital literacy score, Epistemic beliefs (Faith in Intuition), Epistemic beliefs (Faith in Evidence), Epistemic beliefs (Truth is political), Political party (1-7 scale), Affective polarization, Age, Gender (Male = 0, Female = 1), Education, Income, Race/Ethnicity (dummy codes for Black, Hispanic, and Asian identity). Variables were automatically selected to minimize AIC for the overall model. In addition to the general exclusion criteria described above, participants were excluded from regression analyses if they did not provide data for any of the measures included as predictor variables in the regression. For the primary regressions, we also excluded data from those for whom we could not compute an affective polarization score due to neutral partisan preference.

To compare the predictive power of specific coefficients with different DVs, we applied the Seemingly Unrelated Regression (SUR) method, implemented in the R *systemfit* (35) and *car*

packages. This approach estimates two regression models simultaneously and examines whether the $R^2$ value reliably declines if the regression coefficient for a given predictor variable is required to be equal between the two models. For these models, we test all predictors with $p < .10$ for either dependent measure from stepwise regressions. Note that significance testing for each variable in this analysis is computed independently, so we report both uncorrected p values and those obtained after a Bonferroni-Holm correction for multiple comparisons.

In the memory test, for any given candidate and participant, two stimuli were actually "old" and two were actually "new", with the specific assignment of candidate to condition varying based on counterbalancing. The memory test was structured such that only one of the accusation or the accusation control stimuli, and only one of the refutation or refutation control stimuli, would appear in the first half of the memory test, with other stimuli presented in the second half of the memory test. Specific stimuli presented in each half of the test were counterbalanced. For analyses of memory data by condition, we only used data from the first half of the test, to avoid contamination of memory estimates when participants had already seen a matched stimulus earlier in the memory test. Both "Definitely old" and "Probably old" were counted as "old" responses, while "Definitely new" and "Probably new" were counted as "new" responses, and "Not sure" responses were excluded from analysis. Hit rates and false alarm rates were computed for each of the four types of stimuli, with up to 54 trials per condition, and *d'* scores for each stimulus type were calculated with the log-linear correction applied (36). In the separate set of analyses in which candidate impressions were computed based only on trials for which all stimuli were remembered accurately, data from the full memory test data were used, with four stimuli per candidate.

**Experiment 2 – Pre-registered behavioral replication**

**Participants**

As in Experiment 1, we planned to recruit 500 participants using the CloudResearch MTurk Toolkit. Recruitment was stratified by age, with equal samples targeted from five age brackets (18-29, 30-39, 40-49, 50-59, 60+). Due to an error in data collection, an additional set of ~100 participants in the age 30-39 age bracket were recruited, and a slightly smaller sample than anticipated was collected in the age 18-29 age bracket. Participants were paid $7.50 for completing the first part of the study, and an additional $5.00 if they returned for the second part of the study two days later. A total of 561 participants completed part 1; after applying the same exclusion criteria for poor performance as in Experiment 1, 60 participants were excluded, yielding a maximal sample of 501 participants. Only those individuals who met the exclusion criteria in part 1 were invited back for part 2; of these, 381 participants completed part 2.

**Procedure**

Data collection and analysis for Experiment 2 were similar to Experiment 1, following the preregistered plan for Experiment 2 (https://osf.io/ubw6m). We collected the same measures as in Experiment 1 and collected additional individual difference measures on an exploratory basis. The new measures were the Questionnaire of Cognitive and Affective Empathy (37), the 12-item abbreviated Intolerance of Uncertainty (IUS) scale (38), and the 10-item Actively Open-Minded Thinking scale (11, 39). The delayed choice and delayed rating measures were collected both at the end of the first experimental session and the beginning of the second experimental session. The recognition memory test was shifted to the end of the second session and was not administered in the first session. Regression models used the same general approach as in Experiment 1.

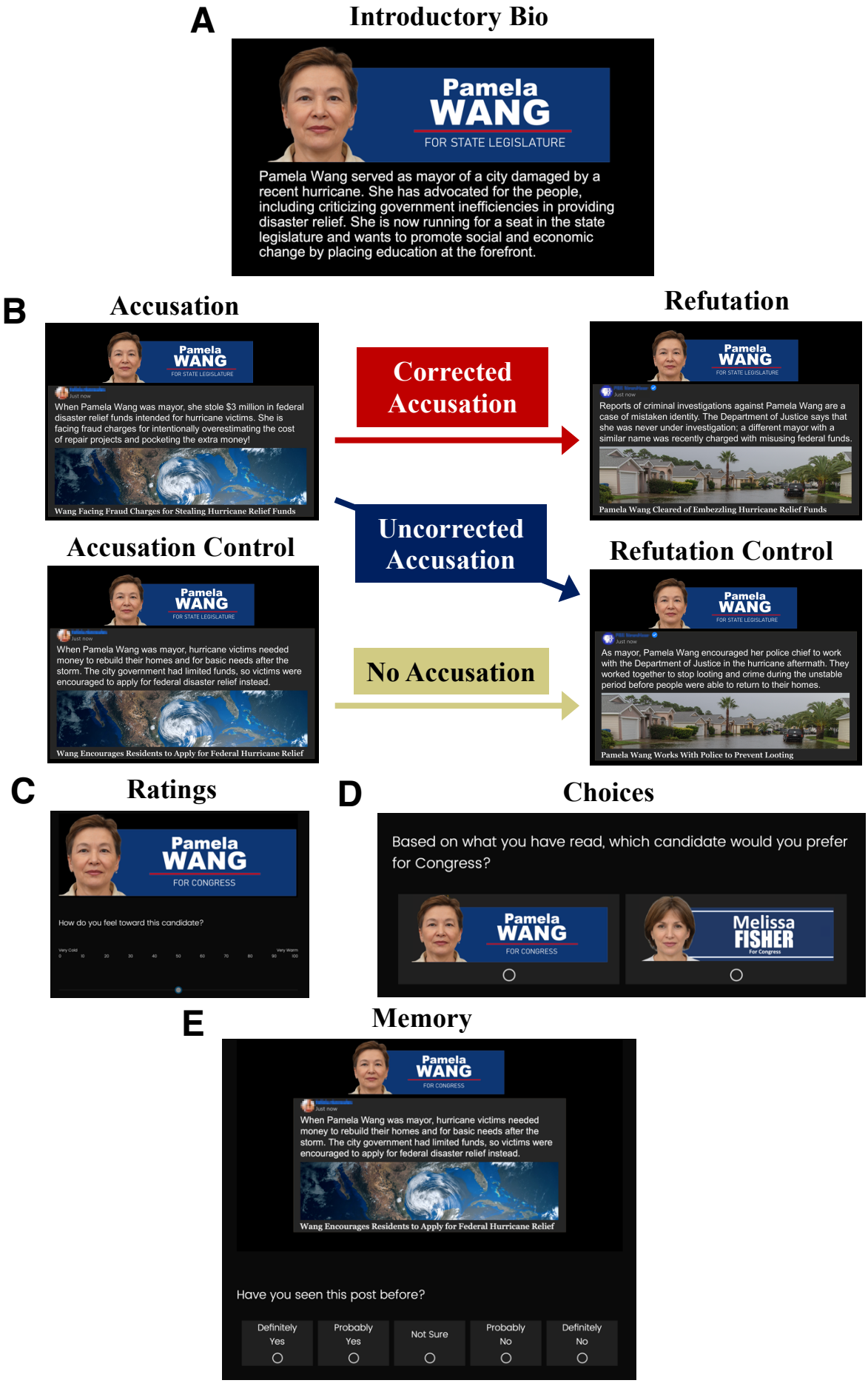



*Figure 1.* Stimulus/task design for (A) introductory bios, (B) core stimuli, (C) ratings (immediate and delayed), (D) delayed choices, (E) memory test.

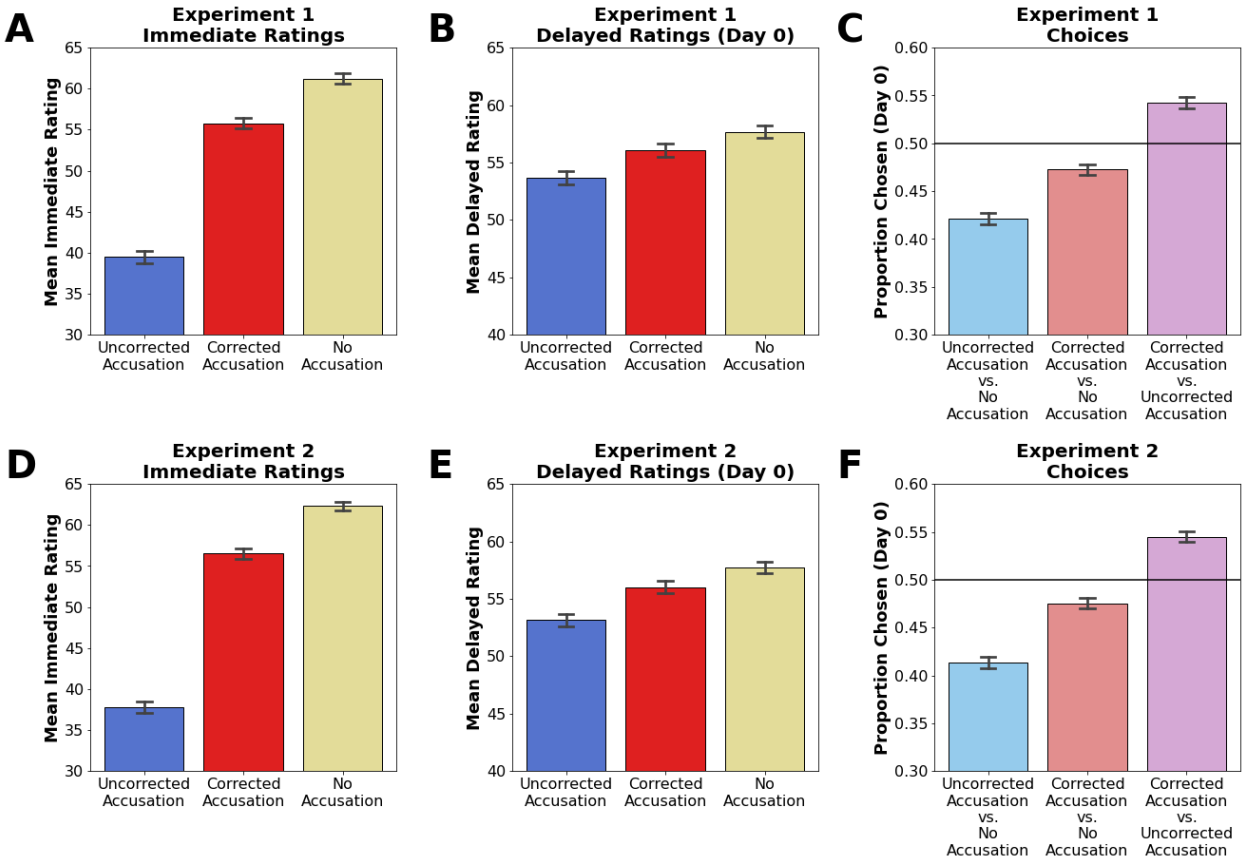



*Figure 2.* Main effects of condition on (A) Experiment 1 immediate ratings, (B) Experiment 1 short-delay ratings, (C) Experiment 1 short-delay choices, (D) Experiment 2 immediate ratings, (E) Experiment 2 short-delay ratings, and (F) Experiment 2 short-delay choices. Effects shown in Experiment 2 constitute a preregistered replication of effects observed in Experiment 1. Error bars represent +/- 1 SE.

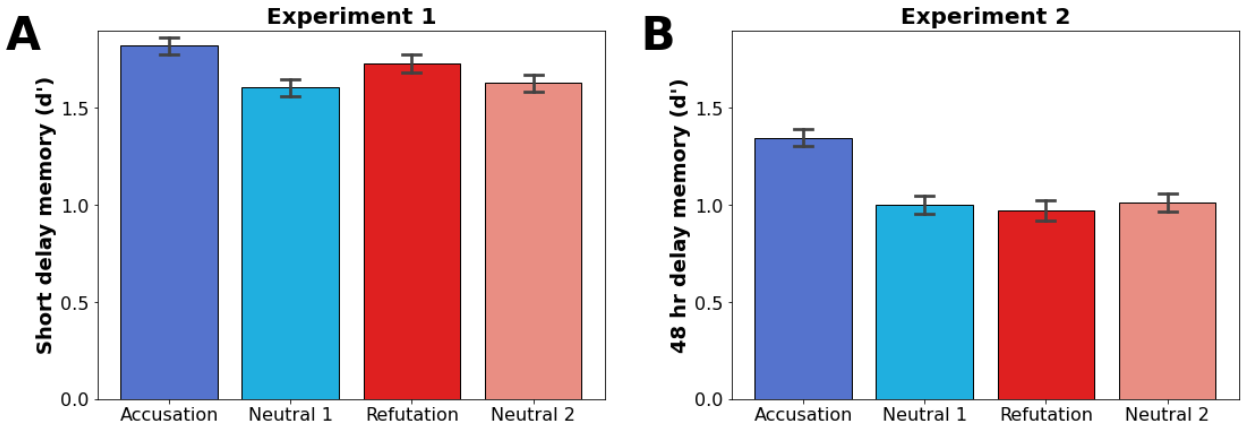



*Figure 3.* Memory by stimulus condition. (A) Experiment 1, memory after a short delay (~30 minutes) (B) Experiment 2, preregistered replication showing memory after a longer (two-day) delay. Error bars represent +/- 1 SE.

**Table 1.** Stepwise regression results showing variables that predict immediate CIEs

**A. Experiment 1 (n = 389)**

| Predictor variable | β | t | p |
|---|---|---|---|
| **Epistemic Beliefs (Faith in Intuition)** | **-0.165** | **-3.260** | **.001**** |
| **Digital Literacy** | **0.123** | **2.411** | **.016*** |
| **Age** | **0.148** | **2.842** | **.005**** |
| **Affective Polarization** | **-0.187** | **-3.842** | **.0001***** |
| Gender (F = 1) | -0.086 | -1.725 | .085 ~ |
| Education | -0.094 | 1.917 | .056 ~ |
| Asian | -0.095 | -1.941 | .053 ~ |

**B. Experiment 2 (n = 425)**

| Predictor variable | β | t | p |
|---|---|---|---|
| **Epistemic Beliefs (Faith in Intuition)** | **-0.162** | **-3.317** | **.001**** |
| Digital Literacy | 0.096 | 1.889 | .06 ~ |
| **Age** | **0.121** | **2.446** | **.015*** |

**Table 2.** SUR regressions comparing effects of predictor variables on immediate CIEs vs. MIST headline accuracy discernment measure.

**A. Experiment 1 (n = 389)**

| Predictor variable | Continued Influence | | MIST | | Difference | | |
|---|---|---|---|---|---|---|---|
| | β | t | β | t | $\chi^2$ | uncorr p | corrected p |
| CRT | 0.048 | 0.92 | 0.119 | **2.69**** | 1.08 | .30 | -- |
| Digital Literacy | 0.109 | **2.06*** | 0.209 | **4.67***** | 2.09 | .15 | -- |
| Epistemic Beliefs (Faith in Intuition) | -0.135 | **-2.35*** | -0.165 | **-3.37***** | 0.15 | .70 | -- |
| Epistemic Beliefs (Evidence) | 0.018 | 0.33 | -0.109 | **2.38*** | 1.65 | .20 | -- |
| Epistemic Beliefs (Truth is Political) | -0.042 | -0.77 | -0.164 | **-3.55***** | 2.91 | .088 | -- |
| **Affective Polarization** | -0.175 | **-3.24**** | 0.085 | 1.86 ~ | **13.49** | **.00023** | **.003**** |
| Belief Superiority | -0.049 | -0.93 | -0.114 | **-2.59*** | 0.92 | .34 | -- |
| **Political Party** | -0.005 | 0.10 | -0.209 | **-4.57***** | **9.17** | **.0025** | **.033*** |
| Age | 0.130 | **2.38*** | 0.100 | **2.15*** | 0.18 | .67 | -- |
| Gender (F = 1) | -0.084 | -1.66 | -0.051 | -1.20 | 0.24 | .62 | -- |
| Education | 0.094 | 1.80 ~ | 0.036 | 0.81 | 0.71 | .40 | -- |
| Income | -0.022 | -0.42 | 0.080 | 1.84 ~ | 2.27 | .13 | -- |
| Race (Black) | -0.014 | -0.28 | -0.191 | **-4.43***** | **7.04** | **.008** | .096 ~ |
| Race (Asian) | -0.092 | -1.84 ~ | -0.002 | -0.05 | 1.89 | .17 | -- |

**B. Experiment 2 (n = 425)**

| Predictor variable | Continued Influence | | MIST | | Difference | | |
|---|---|---|---|---|---|---|---|
| | β | t | β | t | $\chi^2$ | uncorr p | corrected p |
| Digital Literacy | 0.100 | 1.93 ~ | 0.167 | **3.74***** | 0.96 | .33 | -- |
| Epistemic Beliefs (Faith in Intuition) | -0.141 | **-2.68*** | -0.194 | **-4.28***** | 0.58 | .45 | -- |
| Epistemic Beliefs (Truth is Political) | -0.025 | -0.48 | -0.167 | **-3.76***** | **4.37** | **.037*** | .22 |
| **Political Party** | -0.027 | -0.52 | -0.266 | **-5.95***** | **12.19** | **.00048***** | **.0038**** |
| Age | 0.126 | **2.44*** | 0.131 | **2.93**** | 0.004 | .95 | -- |
| Gender (F = 1) | -0.056 | -1.15 | -0.094 | **-2.23*** | 0.34 | .56 | -- |
| Race (Black) | 0.005 | 0.09 | -0.123 | **-2.84**** | 3.70 | .054 ~ | -- |
| Ethnicity (Hispanic) | -0.003 | -0.05 | -0.143 | **-3.43***** | **4.81** | **.028*** | .196 |

**Acknowledgments and Funding Sources**
This work was funded by two grants (in 2020 and 2021) from the Facebook/Meta Research Foundational Integrity research program. Facebook/Meta had no control over the research design, analysis, or publication decisions related to this work. We acknowledge Thomas Hogeboom and Aysha Gsibat for assistance in stimulus development and pilot testing.

## References


1. Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences, 114,* 7313–7318.
2. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359,* 1146-1151.
3. Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1420–1436.
4. Thorson, E. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication, 33,* 460-480.
5. Chan, M. S., Jones, C. R., Hall Jamieson, K., Albarracín, D. (2017). Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science, 28,* 1531-1546.
6. 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.
7. Sirlin, N., Epstein, Z., Arechar, A. A., & Rand, D. G. (2021). Digital literacy is associated with more discerning accuracy judgments but not sharing intentions. *Harvard Kennedy School Misinformation Review.* https://doi.org/10.37016/mr-2020-83 .
8. Jones-Jang, S. M., Mortensen, T., & Liu, J. (2021). Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. *American Behavioral Scientist, 65,* 371-388.

9. Garrett, R. K. & Weeks, B. E. (2017). Epistemic beliefs' role in promoting misperceptions and conspiracist ideation. *PLoS ONE, 12,* e0184733.
10. Van Bavel, J.J., Harris, E., Parnamets, P., Rathje, S., Doell, K. C., & Tucker, J. A. (2021). Political psychology in the digital (mis)information age: A model of news belief and sharing. *Social Issues and Policy Review, 15,* 84-113.
11. Roozenbeek, J., Maertens, R., Herzog, S. M., Geers, M., Kurvers, R., Sultan, M., & van der Linden, S**.** (2022). Susceptibility to misinformation is consistent across question framings and response modes and better explained by myside bias and partisanship than analytical thinking. *Judgment and Decision Making, 17,* 547-573.
12. Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., Petersen, M. B. (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *American Political Science Review, 115*, 999-1015.
13. Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazar, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science, 363,* 374-378.
14. Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances, 5,* eaau4586.
15. Brashier, N. M., & Schacter, D. L. (2020). Aging in an era of fake news. *Current Directions in Psychological Science, 29,* 316-323.
16. Arechar, A. A., Allen, J., Berinsky, A. J., Cole, R., Epstein, Z., Garimella, K., Gully, A., Lu, J. G., Ross, R. M., Stagnaro, M. N., Zhang, Y., Pennycook, G., & Rand, D. G. (2023). Understanding and reducing online misinformation across 16 countries on six continents. *Nature Human Behavior.*

17. Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5,* 296-320.

18. Kensinger, E., & Corkin, S. (2004). Two routes to emotional memory: Distinct neural processes for valence and arousal. *Proceedings of the National Academy of Sciences, 101,* 3310-3315.

19. Zajonc, R.B. (2000). Feeling and thinking: Closing the debate over the independence of affect. In J. P. Forgas (Ed.), *Feeling and thinking: The role of affect in social cognition* (pp. 31–58). Cambridge University Press.

20. Maertens, R., Götz, F. M., Golino, H. F., Roozenbeek, J., Schneider, C. R., Kyrychenko, Y., Kerr, J. R., Stieger, S., McClanahan, W. P., Drabot, K., He, J., & van der Linden, S. (2023). The Misinformation Susceptibility Test (MIST): A psychometrically validated measure of news veracity discernment. *Behavior Research Methods.*

21. Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association, 57,* 348-368.

22. McGaugh, J. (2004). The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annual Review of Neuroscience, 27,* 1-28.

23. Ecker, U.K.H., & Rodricks, A.E. (2020). Do false allegations persist? Retracted misinformation does not continue to influence explicit person impressions. *Journal of Applied Research in Memory and Cognition, 9,* 587–601.

24. Ecker, U.K.H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L.K., Brashier, N., Kendeou, P, Vraga, E.K., & Amazeen, M.A. (2022). The psychological drivers of

misinformation belief and its resistance to correction. *Nature Reviews Psychology, 1,* 13-29.

25. Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications, 5,* 47.
26. Mosleh, M., & Rand, D. G. (2022). Measuring exposure to misinformation from political elites on Twitter. *Nature Communications, 13,* 7144.
27. Lasser, J., Aroyehun, S. T., Simchon, A., Carrella, F., Garcia, D., & Lewandowsky, S. (2022). Social media sharing of low-quality news sources by political elites. *PNAS Nexus, 1,* 1-4.
28. 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,* 23-51.
29. Carstensen, L. L. (2006). The influence of a sense of time on human development. *Science, 312,* 1913-1915.
30. Lempert, K.M., Cohen, M.S., MacNear, K.A., Reckers, F.M., Zaneski, L.A., Wolk, D.A., Kable, J.W. (2022). Aging is associated with maladaptive episodic memory-guided social decision-making. *Proceedings of the National Academy of Sciences, 119,* e2208681119.
31. Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making, 11*, 99–113.
32. Iyengar, S. and Westwood, S.J. (2015), Fear and Loathing across Party Lines: New Evidence on Group Polarization. *American Journal of Political Science, 59,* 690-707.
33. Druckman, J. N., & Levendusky, M. S. (2019). What Do We Measure When We Measure Affective Polarization?. *Public Opinion Quarterly, 83,* 114-122.

34. Toner, K., Leary, M. R., Asher, M. W., & Jongman-Sereno, K. P. (2013). Feeling superior is a bipartisan issue: Extremity (not direction) of political views predicts perceived belief superiority. *Psychological Science*, *24*, 2454–2462.

35. Henningsen, A., & Hamann, J. D. (2007). systemfit: A Package for Estimating Systems of Simultaneous Equations in R. *Journal of Statistical Software*, *23*, 1–40.

36. Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior Research Methods, Instruments, & Computers, 27,* 46-51.

37. Reiners, R. L. E. P., Corcoran, R., Drake, R., Shryane, N. M., & Völlm, B. A. (2011). The QCAE: A questionnaire of cognitive and affective empathy. *Journal of Personality Assessment, 93,* 84-95.

38. Carleton, R. N., Norton, P. A., & Asmundson, G. J. G. (2007). Fearing the unknown: A short version of the Intolerance of Uncertainty Scale. *Journal of Anxiety Disorders, 21,* 105-117.

39. Baron, J., Isler, O., & Yilmaz, O. (2022). Actively open-minded thinking and the political effects of its absence. PsyArXiv. https://psyarxiv.com/g5jhp/

# Supporting Information (SI)

## SI Methods

### Stimulus construction

Face stimuli were selected from AI-generated novel faces available on the Web site https://generated.photos. We aimed to get an even distribution of perceived ages from about 30 to 75 years old. In a pilot study, participants (100 from Amazon MTurk, 44 from the UPenn Psychology participant pool) rated each of 72 faces on attractiveness, warmth, competence, threateningness, likelihood that they would vote for that individual for a political office (e.g., for Congress), and estimated age. Faces were chosen to represent a range of ages and were selected to avoid outliers that were rated either particularly high or particularly low in attractiveness, warmth, competence, threateningness, or likelihood of voting for the person. The final full set of 36 candidate faces is composed of 2/3 male and 1/3 female faces. Of these, 27 are perceived as non-Hispanic white, 3 as Black, 3 as Asian, and 3 as Hispanic, with apparent race initially determined by the study team and later confirmed in pilot data.

Candidate names were generated from a Web-based tool at https://www.name-generator.org.uk/. We began with 99 names, 63 of which could be either White or Black, 18 with Asian surnames, and 18 with Hispanic surnames, as determined by the name generator tool and confirmed by our team. An equal number of names were generated with approximate year of birth in 1989, 1974, and 1959. In a pilot study, 150 participants from Amazon MTurk and 36 participants from the UPenn Psychology participant pool rated a randomly selected sampling of 66 names on the same questions used to evaluate faces. As for the faces, these ratings were used to guide selection of the final set of names so that perceived ages matched between faces and names, and to avoid names that were outliers on other measured features.

Once we had constructed introductory bios, these were presented along with names and faces. These stimuli were pilot-tested on estimated race/ethnicity and judged political party, as well as on the other measures described above (attractiveness, warmth, competence, threateningness, likelihood of voting for that individual for a political office, and estimated age). An initial group of 100 pilot participants were recruited from Amazon MTurk. Bios that were outliers on rating measures were revised and the modified set was then tested on another group of 101 pilot participants from Amazon MTurk. Ratings of judged political party from this final sample were used for the analysis in the main text examining how alignment between candidate political orientation and political orientation affected candidate ratings.

Finally, we added the accusations, refutations, and matched control stimuli to the paradigm. In this phase of pilot testing, we examined the difference between feeling thermometer ratings made immediately after reading the two critical story stimuli (post-story) and an earlier feeling thermometer rating made after only the introductory bio (pre-story) to determine whether each item was behaving as expected.

CIEs were apparent in the aggregate even in the first full version of our stimulus set (see SI Results). Still, we made further modifications to the stimuli to ensure that for each item, post-story ratings in the Uncorrected Accusation condition would be much lower than pre-story ratings, post-story ratings in the No Accusation condition would be unchanged from pre-story ratings, and post-story ratings in the Corrected Accusation condition would decline relative to pre-story ratings but less than declines in the Uncorrected Accusation condition. The 36-item stimulus set used in pilot testing, as well as in our forthcoming neuroimaging study, included 3 additional types of accusations (illegal campaign contributions, sexual misconduct, and murder), and a fourth political office (state legislature). We ran 7 rounds of pilot testing, with edits made

to the stimuli after each round. Participant counts were as follows: Round 1 included a sample of 62 participants from the UPenn Psychology subject pool in addition to the 150 participants from CloudResearch MTurk Toolkit whose data are described below; Round 2 included 81 participants from UPenn Psychology subject pool; Round 3 included 185 participants from CloudResearch MTurk Toolkit; Round 4 included 145 participants from CloudResearch MTurk Toolkit; Round 5 included 150 participants from CloudResearch MTurk Toolkit; Round 6 included 151 participants from CloudResearch MTurk Toolkit; Round 7 included 152 participants from CloudResearch MTurk Toolkit.

## SI Results

### Accusation and Correction Effects

Strong accusation effects were apparent on immediate ratings in Experiment 1, $t(457) = -28.80$, $p < .001$, $d = -1.35$. Note that accusation effects, like CIEs, are computed such that negative scores indicate larger effects. Accusation effects remained present on ratings measured after a short delay, $t(457) = -11.80$, $p < .001$, $d = -.55$, and were also present on choices after a short delay, $t(457) = -13.14$, $p < .001$, $d = -.61$. In preregistered analyses in Experiment 2, the accusation effect replicated on immediate ratings, $t(500) = -32.90$, $p < .001$, $d = -1.47$, on ratings made after a short delay, $t(500) = -12.38$, $p < .001$, $d = -.55$, and on choices made after a short delay, $t(500) = -14.56$, $p < .001$, $d = -.65$. Finally, accusation effects remained after a 48-hour delay in Experiment 2 on ratings, $t(423) = -12.84$, $p < .001$, $d = -.62$, and on choices, $t(423) = -12.98$, $p < .001$, $d = -.63$. Accusation effects did not differ in Experiment 2 between short delay and long delay measures for ratings, $t(423) = 1.34$, $p = .18$, $d = .07$. There was a difference between short delay and long delay choice measures, however, $t(423) = -2.42$, $p = .016$, $d = -.12$,

as candidates with an uncorrected accusation were slightly more likely to be chosen after a long delay ($M$ = 42.31%, $SE$ = 0.59%) versus after a short delay ($M$ = 41.35%, $SE$ = 0.59%).

Correction effects, in which refutations improve impressions relative to an uncorrected accusation, were strongly apparent in immediate ratings in Experiment 1, $t(457) = 24.43$, $p < .001$, $d = 1.14$. Here, positive values indicate a larger increase in preference as a result of the correction. These effects remained significant after a short delay for both ratings, $t(457) = 7.13$, $p < .001$, $d = .33$, and choices, $t(457) = 7.58$, $p < .001$, $d = .35$. Preregistered analyses in Experiment 2 show that correction effects on immediate ratings replicated, $t(500) = 27.76$, $p < .001$, $d = 1.24$, as did correction effects measured at a short delay for both ratings, $t(500) = 8.97$, $p < .001$, $d = .40$, and choices, $t(500) = 8.27$, $p < .001$, $d = .37$. Finally, correction effects remained significant after a 48-hour delay in Experiment 2 for both ratings, $t(423) = 8.91$, $p < .001$, $d = .43$, and choices, $t(423) = 6.77$, $p < .001$, $d = .33$. There were no differences between short-delay and long-delay correction effects, for ratings, $|t|(423) < 1$, $d = -.04$, or for choices, $t(423) < 1$, $d = .03$.

**Accusation and correction effects conditionalized on successful memory**

As was reported in the main text regarding CIEs, accusation effects remained when limiting analyses to trials in which all four stimuli for that candidate were subsequently remembered correctly as having been seen or not. Accusation effects on immediate ratings remained robust in Experiment 1, $t(312) = -21.37$, $p < .001$, $d = -1.21$, and in Experiment 2, $t(155) = -17.87$, $p < .001$, $d = -1.43$. Accusation effects on ratings also remained robust after a short delay in Experiment 1, $t(312) = -8.95$, $p < .001$, $d = -.51$, after a short delay in Experiment 2, $t(155) = -5.36$, $p < .001$, $d = -.43$, and after a long delay in Experiment 2, $t(155) = -6.66$, $p < .001$, $d = -.53$. Finally, accusation effects remained present on the choice task, at a short delay in

Experiment 1, $t(188) = -8.09$, $p < .001$, $d = -.59$, at a short delay in Experiment 2, $t(53) = -3.32$, $p = .002$, $d = -.45$, and at a long delay in Experiment 2, $t(53) = -3.89$, $p < .001$, $d = -.53$.

Similarly, the benefits of corrections remained strong among candidates for which all stimuli were subsequently remembered. Correction effects were apparent on immediate ratings in Experiment 1, $t(305) = 17.49$, $p < .001$, $d = 1.00$, and in Experiment 2, $t(149) = 13.71$, $p < .001$, $d = 1.12$. Correction effects were also reliable on ratings made at a short delay in Experiment 1, $t(305) = 3.53$, $p < .001$, $d = .20$, at a short delay in Experiment 2, $t(149) = 3.08$, $p = .002$, $d = .25$, and at a long delay in Experiment 2, $t(149) = 3.44$, $p < .001$, $d = .28$. Finally, correction effects on candidate choices were apparent at a short delay in Experiment 1, $t(193) = 5.24$, $p < .001$, $d = .38$, at a short delay in Experiment 2, $t(49) = 3.25$, $p = .002$, $d = .46$, and at a long delay in Experiment 2, $t(49) = 2.58$, $p = .013$, $d = .36$.

**Relationship between perceived candidate ideology, participant ideology, and CIEs**

We examined whether CIEs differ based on whether the mock candidate is perceived to align with a participant's ideology. Measures of perceived candidate ideology were obtained from a separate pilot sample that only saw the bios. These perceptions of candidate ideology and participants' self-reported ideological orientation were regressed onto candidate ratings in exploratory linear mixed effects analyses, controlling for participants' initial ratings made after reading the candidate bios (see Supplemental Table S10 for the full regression model). In both experiments, ratings in the no accusation condition were higher when candidate ideology and participant ideology were more aligned, as evidenced by positive two-way interaction effects (Experiment 1: $\beta = 0.059$, $t = 4.75$, $p < .001$; Experiment 2: $\beta = 0.082$, $t = 6.74$, $p < .001$). In Experiment 2, three-way interactions indicate that this effect was reliably reduced for candidates in the corrected accusation ($\beta = -0.022$, $t = -2.22$, $p = .027$) and uncorrected accusation

(β = -0.037, $t$ = -3.78, $p$ < .001) conditions. In Experiment 1, these effects were not significant, but the trends were towards a reduced effect of ideological alignment in both the corrected accusation (β = -0.012, $t$ = -1.20, $p$ = .23) and uncorrected accusation (β = -0.018, $t$ = -1.76, $p$ = .078) conditions. We thus conclude that ideological alignment is not protective against CIEs and may in fact worsen them.

**Relationship between memory and CIEs**

In exploratory follow-up analyses, we find evidence potentially linking the initial processing of accusations with increased persistence of refuted accusations in decision making. Specifically, in Experiment 2, better memory two days later for accusation stimuli, relative to matched neutral stimuli, correlated with greater immediate CIEs, $r$(403) = -.16, $p$ < .001 (see Supplemental Figure S2). An analogous analysis showed no reliable effect for refutation stimuli, $r$(403) = -.07, $p$ = .19. However, the difference between these effects for accusation and refutation stimuli was not significant, $p$ = .18. In Experiment 1, when memory was measured at a short delay, there was no relationship between CIEs and memory for accusation stimuli, $r$(436) = .01, $p$ = .85, or for refutation stimuli, $r$(436) = .02, $p$ = .68. The difference between experiments is likely due to effects of selective consolidation that were apparent by the time of the memory test in Experiment 2, two days after encoding, but had not yet emerged at the time of the Experiment 1 memory test, similar to what we observe in the analysis of memory by condition.

Note that because memory for accusations two days later correlated with CIEs that were computed from immediate ratings, a causal link from memory to ratings is temporally implausible. Instead, we interpret this relationship as tentative evidence for a common mechanism by which prioritized initial processing of accusations, perhaps due to an increase in emotional arousal, strengthens both memory and CIEs. The case for this common mechanism

would be stronger if the relationship between accusations and CIEs were significantly stronger than the relationship between refutations and CIEs. As it is, an alternate explanation, that individuals who are more engaged in the task show both better memory for impactful stimuli at a two-day delay and larger CIEs, cannot be ruled out. Future work in which emotional arousal is measured more directly than in the present studies would be better positioned to address these issues.

**Initial Pilot data**

In order to demonstrate that CIEs were present with these stimuli prior to any modifications, we focus on the 150 participants recruited from CloudResearch MTurk for pilot Experiment 1 as a comparable sample to that in the final experiments. Here, we found that immediate ratings for corrected accusation stimuli ($M$ = 52.9, $SE$ = 1.22) were lower than immediate ratings for no accusation stimuli ($M$ = 62.2, $SE$ = 1.01), indicating a significant continued influence effect, $t(150) = -9.72$, $p < .001$, $d = -.79$. Uncorrected accusation stimuli ($M$ = 39.5, $SE$ = 1.19) were also rated lower than no accusation stimuli, indicating a significant accusation effect, $t(150) = -17.68$, $p < .001$, $d = -1.44$. Finally, corrected accusation stimuli were rated higher than uncorrected accusation stimuli, $t(150) = 12.03$, $p < .001$, $d = .98$, indicating a significant correction effect. Similarly, on a short-delay choice task, we found evidence for a significant CIE on choices between corrected accusation and no accusation stimuli, $M$ = 46.7%, $SE$ = 0.8%, $t(150) = -4.28$, $p < .001$, $d = -.35$ . We also found evidence for a significant accusation effect on choices between uncorrected accusation and no accusation stimuli, $M$ = 45.5%, $SE$ = 0.8%, $t(150) = -5.76$, $p < .001$, $d = -.47$. We did not find a correction effect on choices between corrected accusation and uncorrected accusation stimuli, $M$ = 50.7%, $SE$ = 0.8%, $t(150) < 1$, $d = .07$.

**Pilot data comparing short vs. long delay assessments**

While in most aspects our pilot data were highly similar to the main study with respect to main effects, one set of findings from the pilot data requires more detailed explanation, as these data inspired a prediction in our preregistration that was not confirmed. Specifically, we expected to find that the degree to which corrections ameliorated the impact of accusations on choices would be reduced after a longer delay. This expectation was based largely on data from rounds 5 and 6 of pilot testing, in which assessments after a 48-hour delay were included.

Both round 5 and round 6 of the pilot study included a choice task at both short-delay (5-10 min) and long-delay (2 days later) time points. We report those data combined across these two experiments. We found that CIEs on the choice task grew from the short-delay assessment to the long-delay assessment, $t(222) = 2.95$, $p = .004$, $d = .20$. Specifically, the proportion of trials between a corrected accusation and a no accusation candidate for which the corrected accusation candidate was chosen dropped from a mean of 48.1% (SE = 0.6%) to a mean of 46.4% (SE = 0.7%). At the same time, accusation effects on the choice task were weaker after a long delay, $t(222) = -2.30$, $p = .022$, $d = -.16$, as choices of uncorrected accusation candidates over no accusation candidates increased from a mean of 44.0% (SE = 0.7%) at a short delay to a mean of 45.4% (SE = 0.7%) after a long delay. The magnitude of correction effects on the choice task did not reliably change, $t(222) = 1.37$, $p = .17$, $d = .09$, though there was a trend for reduced correction effects after a longer delay, from a mean of 53.1% (SE = 0.7%) at a short delay to a mean of 52.2% (SE = 0.7%) at a long delay.

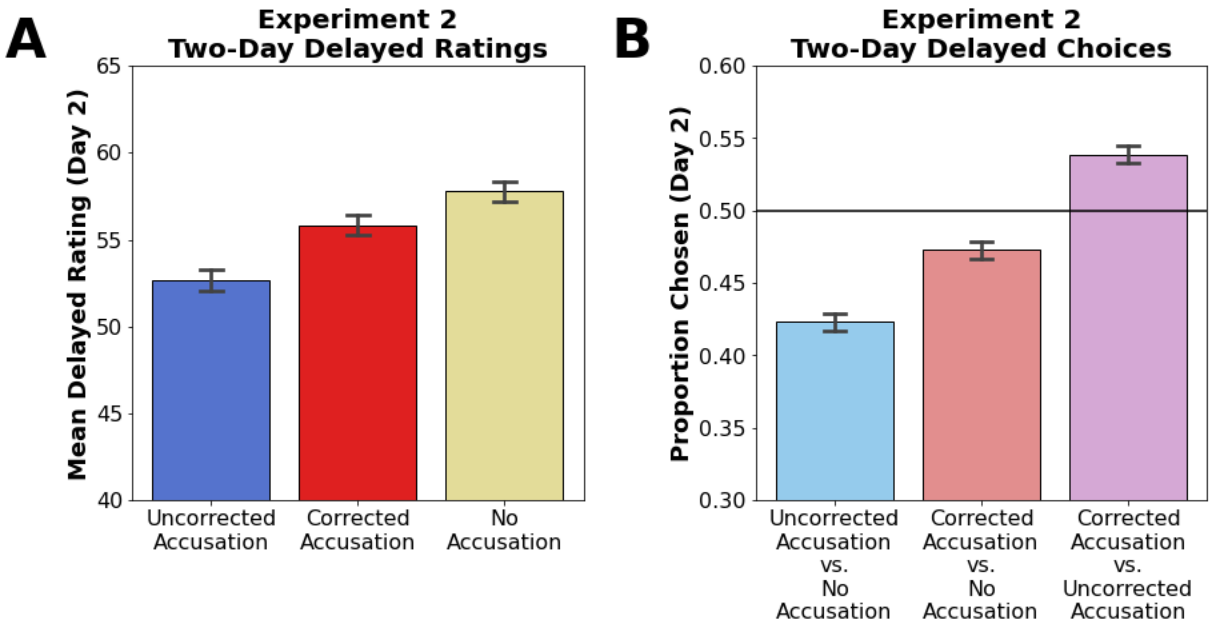
Short-delay feeling thermometer ratings were not included in round 5, so differences between short-delay and long-delay ratings could only be examined on data from round 6. Here, CIEs did not reliably differ from the short-delay assessment (mean ratings of 55.3 for corrected

accusation stimuli vs. 57.2 for no accusation stimuli) to the long-delay assessment (mean ratings of 54.1 for corrected accusation stimuli vs. 56.1 for no accusation stimuli), $t(94) < 1$, $d = .04$. At the same time, correction effects reliably declined from the short delay assessment (mean ratings of 55.3 for corrected accusation stimuli vs. 53.4 for uncorrected accusation stimuli) to the long delay assessment (mean ratings of 54.1 for corrected accusation stimuli vs. 53.7 for uncorrected accusation stimuli), $t(94) = 3.02$, $p = .003$, $d = .31$. Accusation effects also declined from the short delay assessment (mean ratings of 53.4 for uncorrected accusation stimuli vs. 57.2 for no accusation stimuli) to the long delay assessment (mean ratings of 53.7 for uncorrected accusation stimuli vs. 56.1 for uncorrected accusation stimuli), $t(94) = -2.33$, $p = .022$, $d = -.24$.

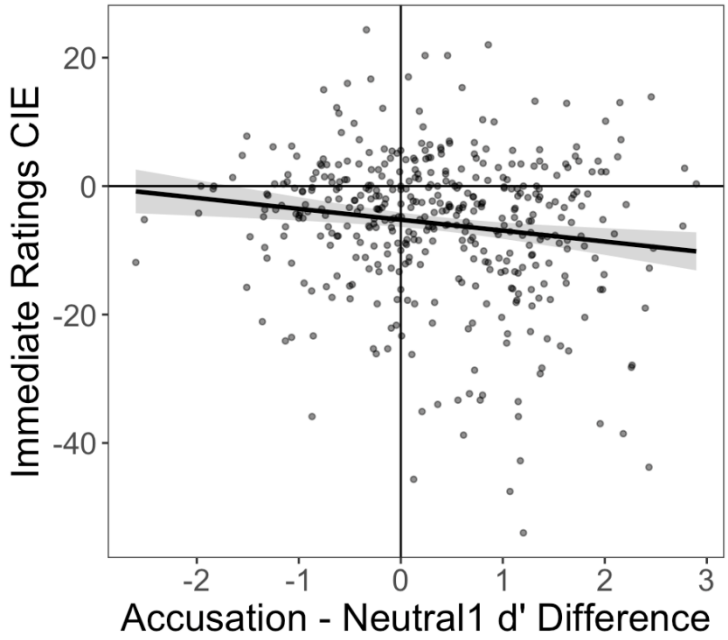
Thus, our pilot work suggested that corrections were likely to lose effectiveness over time, leading to greater CIEs and/or reduced correction effects on long-delay vs. short-delay measures. It is unclear why we did not find such a result in the main studies. One difference between the pilot studies and the main studies was the reduction in the number of candidate stimuli from 36 to 27 items. Additionally, the delay interval before the short-delay assessments was somewhat longer in the main studies compared to the pilot studies, because additional questionnaire measures were placed after the initial presentations of the core stimuli and immediate ratings but before the delayed rating phase in the main study. This difference may have made the short-delay and long-delay ratings more similar to each other in terms of cognitive processing. It is of course also possible that the original finding was a Type I error, or that the failure to replicate was a Type II error. Further work will be necessary to examine these questions.

**Digital literacy measure**

The digital literacy measure was constructed using an initial set of 8 questions. One question was taken directly from Sirlin et al., 2021 (ref. 7 in the main text), while others were written by our team. We ran a pilot test on this first version using 78 participants recruited from CloudResearch MTurk Toolkit. Of the eight questions, one-sample t-tests showed that five questions yielded accuracy well above chance, while three were at chance-level accuracy. For the second pilot version, these three questions were modified, and the modified scale was tested on another pilot sample of 80 participants from CloudResearch MTurk Toolkit. Here, all stimuli were above chance accuracy. Additionally, we examined Cronbach's alpha for the second pilot version, which for the full 8-item scale was .66. There were two questions for which Cronbach's alpha increased if the item was deleted. Thus, we removed those two items to yield the final six-item scale described in the Supplemental Appendix. Cronbach's alpha among the pilot sample for these 6 items was .72, which we determined to be acceptable reliability.




*Supplemental Figure S1.* Aggregate behavioral effects by condition in Experiment 2 after a long (2-day) delay on (A) ratings and (B) choices. Error bars represent +/- 1 SE.




*Supplemental Figure S2.* Relationship between CIEs computed from immediate ratings and the memory benefit for accusation stimuli relative to neutral stimuli in Experiment 2. Shaded regions represent 95% confidence interval.

Supplemental Table S1. Stepwise regression results showing variables that predict immediate accusation effects

**A. Experiment 1 (n = 389)**

| Predictor variable | β | t | p |
|---|---|---|---|
| **Affective Polarization** | **-0.234** | **-4.749** | **< .0001*** ** |
| Education | 0.092 | 1.856 | .064 ~ |
| Gender (F = 1) | -0.082 | -1.666 | .097 ~ |

**B. Experiment 2 (n = 425)**

| Predictor variable | β | t | p |
|---|---|---|---|
| **Epistemic Beliefs (Faith in Intuition)** | **-0.190** | **-3.700** | **.0002*** ** |
| **Epistemic Beliefs (Need for Evidence)** | **-0.144** | **-2.770** | **.006** ** |
| **CRT** | **-0.149** | **-3.121** | **.002** ** |
| **Asian** | **-0.099** | **-2.092** | **.037*** |
| Gender (F = 1) | -0.087 | -1.783 | .075 ~ |
| Education | 0.083 | 1.709 | .088 ~ |

Supplemental Table S2. Stepwise regression results showing variables that predict immediate correction effects

**A. Experiment 1 (n = 389)**

| Predictor variable | β | t | p |
|---|---|---|---|
| **Epistemic Beliefs (Truth is Political)** | **-0.120** | **-2.346** | **.020*** |
| **Age** | **0.118** | **2.312** | **.021*** |
| Digital Literacy | 0.095 | 1.862 | .063 ~ |
| Affective Polarization | 0.093 | 1.820 | .070 ~ |
| Epistemic Beliefs (Need for Evidence) | 0.085 | 1.649 | .10 |

**B. Experiment 2 (n = 425)**

| Predictor variable | β | t | p |
|---|---|---|---|
| **Epistemic Beliefs (Need for Evidence)** | **0.185** | **3.592** | **.0004***** |
| **CRT** | **0.156** | **3.252** | **.001**** |
| Epistemic Beliefs (Faith in Intuition) | 0.087 | 1.693 | .091 ~ |
| Gender (F = 1) | 0.077 | 1.570 | .117 |

Supplemental Table S3. Stepwise regression results showing variables that predict delayed CIEs

**A. Experiment 1 – Short delay (n = 389)**

| Predictor variable | β | t | p |
|---|---|---|---|
| Education | 0.064 | 1.467 | .143 |
| Epistemic Beliefs (Truth is Political) | -0.064 | -1.445 | .149 |

**B. Experiment 2 – Short delay (n = 425)**

| Predictor variable | β | t | p |
|---|---|---|---|
| **Affective Polarization** | **0.092** | **2.138** | **.033*** |
| **Gender (F = 1)** | **-0.124** | **-2.875** | **.004**** |
| CRT | -0.079 | -1.822 | .069 ~ |

**C. Experiment 2 – Long delay (n = 361)**

| Predictor variable | β | t | p |
|---|---|---|---|
| **Epistemic Beliefs (Truth is Political)** | **-0.110** | **-2.335** | **.020*** |
| Gender (F = 1) | -0.088 | -1.884 | .060 ~ |
| CRT | -0.083 | -1.739 | .083 ~ |
| Digital Literacy | -0.070 | -1.471 | .142 |
| Affective Polarization | 0.076 | 1.573 | .117 |

Supplemental Table S4. SUR regression results showing differences between predictors of immediate and delayed CIEs

**A. Experiment 1 – Short delay (n = 389)**

| Predictor variable | Immediate | | Delayed | | Difference | | |
|---|---|---|---|---|---|---|---|
| | β | t | β | t | $\chi^2$ | uncorr p | corrected p |
| Digital literacy | **0.123** | **2.41*** | -0.032 | -0.67 | **4.95** | **.026** | .156 |
| Epistemic Beliefs (Faith in Intuition) | **-0.165** | **-3.26**** | -0.046 | -0.99 | 2.97 | .085 | -- |
| Age | **0.148** | **2.84**** | 0.001 | 0.01 | **4.32** | **.038** | .19 |
| Gender (F = 1) | -0.086 | -1.72 | -0.055 | -1.19 | 0.21 | .65 | -- |
| Affective Polarization | -0.187 | **-3.84**** | 0.010 | 0.22 | **8.83** | **.003** | **.021*** |
| Education | 0.094 | 1.92 | 0.060 | 1.32 | 0.26 | .61 | -- |
| Asian | -0.095 | -1.94 | -0.008 | -0.18 | 1.71 | .19 | -- |

**B. Experiment 2 – Short delay (n = 425)**

| Predictor variable | Immediate | | Delayed | | Difference | | |
|---|---|---|---|---|---|---|---|
| | β | t | β | t | $\chi^2$ | uncorr p | corrected p |
| Digital literacy | **0.109** | **2.09*** | -0.048 | -1.02 | **4.99** | **.025** | .125 |
| Epistemic Beliefs (Faith in Intuition) | **-0.153** | **-3.05**** | 0.027 | 0.60 | **7.13** | **.0076** | **.046*** |
| Age | **0.130** | **2.60**** | 0.029 | 0.65 | 2.23 | .14 | -- |
| Gender (F = 1) | -0.053 | -1.09 | **-0.126** | **-2.88**** | 1.26 | .26 | -- |
| CRT | -0.030 | -0.60 | -0.064 | -1.44 | 0.27 | .60 | -- |
| Affective Polarization | -0.038 | -0.78 | 0.089 | **2.04*** | 3.80 | .051 | -- |

**C. Experiment 2 – Long delay (n = 361)**

| Predictor variable | Immediate | | Delayed | | Difference | | |
|---|---|---|---|---|---|---|---|
| | β | t | β | t | $\chi^2$ | uncorr p | corrected p |
| **Digital literacy** | **0.114** | **2.00*** | -0.057 | -1.12 | **4.98** | **.026** | .156 |
| Epistemic Beliefs (Faith in Intuition) | **-0.127** | **-2.25*** | -0.023 | -0.46 | 1.88 | .17 | -- |
| Age | 0.100 | 1.85 ~ | 0.056 | 1.16 | 0.37 | .55 | -- |
| Gender (F = 1) | -0.060 | -1.14 | 0.086 | -1.80 ~ | 0.12 | .73 | -- |
| Epistemic Beliefs (Truth is Political) | -0.054 | -1.00 | -0.099 | **-2.02*** | 0.37 | .54 | -- |
| CRT | -0.038 | -0.71 | -0.092 | -1.93 ~ | 0.58 | .45 | -- |

Supplemental Table S5. Stepwise regression results showing variables that predict delayed accusation effects

**A. Experiment 1 – Short delay (n = 389)**

| Predictor variable | β | t | p |
|---|---|---|---|
| **Digital Literacy** | **-0.092** | **-2.079** | **.038*** |
| **Gender (F = 1)** | **-0.088** | **-1.978** | **.049*** |

**B. Experiment 2 – Short delay (n = 425)**

| Predictor variable | β | t | p |
|---|---|---|---|
| **Digital Literacy** | **-0.100** | **-2.295** | **.022*** |
| CRT | -0.080 | -1.816 | .070 ~ |
| Gender (F = 1) | -0.076 | -1.779 | .076 ~ |
| Asian | -0.068 | -1.587 | .113 |

**C. Experiment 2 – Long delay (n = 361)**

| Predictor variable | β | t | p |
|---|---|---|---|
| **Digital Literacy** | **-0.123** | **-2.558** | **.011*** |
| Asian | -0.069 | -1.497 | .135 |
| Epistemic Beliefs (Truth is Political) | -0.074 | -1.462 | .145 |
| Epistemic Beliefs (Belief in Evidence) | -0.074 | -1.439 | .151 |
| Gender (F = 1) | -0.080 | -1.616 | .107 |

Supplemental Table S6. SUR regression results showing differences between predictors of immediate and delayed accusation effects

**A. Experiment 1 – Short delay (n = 389)**

| Predictor variable | Immediate | | Delayed | | Difference | | |
|---|---|---|---|---|---|---|---|
| | β | t | β | t | $\chi^2$ | uncorr p | corrected p |
| Digital Literacy | 0.004 | 0.09 | -0.090 | **-2.02*** | 2.01 | .16 | -- |
| **Affective Polarization** | **-0.234** | **-4.74***** | 0.014 | 0.31 | **13.92** | **.0002** | **.0008***** |
| Gender (F = 1) | -0.082 | -1.67 | -0.085 | -1.91 ~ | 0 | .97 | -- |
| Education | 0.092 | 1.86 ~ | 0.054 | 1.21 | 0.33 | .57 | -- |

**B. Experiment 2 – Short delay (n = 425)**

| Predictor variable | Immediate | | Delayed | | Difference | | |
|---|---|---|---|---|---|---|---|
| | β | t | β | t | $\chi^2$ | uncorr p | corrected p |
| Digital literacy | 0.023 | 0.46 | **-0.097** | **-2.16*** | 3.22 | .073 ~ | -- |
| **Epistemic Beliefs (Faith in Intuition)** | **-0.185** | **-3.54***** | -0.001 | -0.03 | **6.77** | **.009** | .063 ~ |
| Epistemic Beliefs (Need for Evidence) | **-0.145** | **-2.78**** | -0.036 | -0.76 | 2.39 | .12 | -- |
| Gender (F = 1) | -0.089 | -1.82 ~ | -0.082 | -1.86 ~ | 0.01 | .92 | -- |
| CRT | **-0.153** | **-3.15**** | -0.079 | -1.77 ~ | 1.29 | .26 | -- |
| Asian | **-0.099** | **-2.09*** | -0.068 | -1.58 | 0.23 | .63 | -- |
| Education | 0.084 | 1.72 ~ | 0.010 | 0.22 | 1.26 | .26 | -- |

**C. Experiment 2 – Long delay (n = 361)**

| Predictor variable | Immediate | | Delayed | | Difference | | |
|---|---|---|---|---|---|---|---|
| | β | t | β | t | $\chi^2$ | uncorr p | corrected p |
| **Digital literacy** | 0.062 | 1.16 | **-0.120** | **-2.41*** | **6.25** | **.012** | .084 ~ |
| Epistemic Beliefs (Faith in Intuition) | **-0.149** | **-2.61**** | -0.064 | -1.20 | 1.18 | .28 | -- |
| Epistemic Beliefs (Need for Evidence) | **-0.126** | **-2.23*** | -0.084 | -1.58 | 0.30 | .58 | -- |
| Gender (F = 1) | 0.083 | -1.56 | -0.068 | -1.35 | 0.05 | .83 | -- |
| CRT | -0.129 | **-2.47*** | -0.048 | -0.97 | 1.29 | .26 | -- |
| Asian | -0.090 | -1.82 ~ | -0.078 | -1.69 | 0.03 | .86 | -- |
| Education | 0.099 | 1.86 ~ | 0.057 | 1.14 | 0.34 | .56 | -- |

Supplemental Table S7. Stepwise regression results showing variables that predict delayed correction effects

**A. Experiment 1 – Short delay (n = 389)**

| Predictor variable | β | t | p |
|---|---|---|---|
| **Epistemic Beliefs (Faith in Intuition)** | **-0.099** | **-2.176** | **.030*** |

**B. Experiment 2 – Short delay (n = 425)**

| Predictor variable | β | t | p |
|---|---|---|---|
| Digital Literacy | 0.070 | 1.632 | .103 |

**C. Experiment 2 – Long delay (n = 361)**

(No variables in model)

Supplemental Table S8. SUR regression results showing differences between predictors of immediate and delayed correction effects

**A. Experiment 1 – Short delay (n = 389)**

| Predictor variable | Immediate | | Delayed | | Difference | | |
|---|---|---|---|---|---|---|---|
| | β | t | β | t | $\chi^2$ | uncorr p | corrected p |
| Epistemic Beliefs (Truth is Political) | **-0.126** | **-2.33**** | -0.038 | -0.76 | 1.46 | .23 | -- |
| **Age** | **0.118** | **2.30*** | -0.051 | -1.08 | **5.86** | **.015**** | **.075 ~** |
| Digital Literacy | 0.096 | 1.84 ~ | 0.007 | 0.15 | 1.58 | .21 | -- |
| Epistemic Beliefs (Faith in Intuition) | -0.031 | -0.56 | -0.083 | -1.65 | 0.49 | .48 | -- |
| **Affective Polarization** | **0.109** | **2.17*** | -0.029 | -0.64 | **4.11** | **.043*** | .17 |

**B. Experiment 2 – Short delay (n = 425)**

| Predictor variable | Immediate | | Delayed | | Difference | | |
|---|---|---|---|---|---|---|---|
| | β | t | β | t | $\chi^2$ | uncorr p | corrected p |
| **Epistemic Beliefs (Need for Evidence)** | **0.172** | **3.38***** | 0.020 | 0.43 | **4.90** | **.027*** | .081 ~ |
| CRT | **0.152** | **3.16**** | 0.047 | 1.08 | 2.61 | .106 | -- |
| Epistemic Beliefs (Faith in Intuition) | 0.095 | 1.84 ~ | 0.004 | 0.08 | 1.72 | .19 | -- |

**C. Experiment 2 – Long delay (n = 361)**

| Predictor variable | Immediate | | Delayed | | Difference | | |
|---|---|---|---|---|---|---|---|
| | β | t | β | t | $\chi^2$ | uncorr p | corrected p |
| **Epistemic Beliefs (Need for Evidence)** | **0.166** | **2.97**** | 0.005 | 0.10 | 4.44 | **.035*** | .105 |
| CRT | **0.108** | **2.08*** | -0.003 | -0.06 | 2.45 | .12 | -- |
| Epistemic Beliefs (Faith in Intuition) | 0.079 | 1.40 | -0.020 | -0.39 | 1.67 | .20 | -- |

Supplemental Table S9. Mixed effects models examining effects of perceived candidate ideology, participant ideology, stimulus condition, and affective polarization on immediate ratings.

Model specification: Immediate Post-Story Rating ~ Initial Pre-Story Rating + Participant Ideology * Candidate Ideology * Affective Polarization * Corrected Accusation + Participant Ideology * Candidate_Ideology * Affective_Polarization * Uncorrected Accusation + Counterbalance1 + Counterbalance2 + (1|Subject)

**Experiment 1 (n = 401)**

| **Predictor** | **β** | **t** | **p** |
|---|---|---|---|
| **Initial Rating** | **0.170** | **17.736** | **< 0.0001*** ** |
| **Participant Political Orientation** | **-0.061** | **-2.368** | **0.018*** |
| Candidate Political Orientation | -0.011 | -0.809 | 0.419 |
| Affective Polarization Score | 0.034 | 1.251 | 0.212 |
| **Corrected Accusation (vs. Neutral) Stimulus** | **-0.113** | **-12.788** | **< 0.0001*** ** |
| **Uncorrected Accusation (vs. Neutral) Stimulus** | **-0.448** | **-50.663** | **< 0.0001*** ** |
| Counterbalance1 | 0.015 | 0.557 | 0.578 |
| Counterbalance2 | -0.002 | -0.077 | 0.938 |
| **Participant Political Orientation x Candidate Political Orientation** | **0.055** | **3.956** | **< 0.0001*** ** |
| Participant Political Orientation x Affective Polarization Score | -0.007 | -0.252 | 0.801 |
| Candidate Political Orientation x Affective Polarization Score | 0.008 | 0.621 | 0.534 |
| Participant Political Orientation x Corrected Accusation (vs. Neutral) Stimulus | -0.018 | -1.592 | 0.111 |
| Candidate Political Orientation x Corrected Accusation (vs. Neutral) Stimulus | -0.014 | -1.296 | 0.195 |
| **Affective Polarization Score x Corrected Accusation (vs. Neutral) Stimulus** | **-0.040** | **-3.534** | **0.0004*** ** |
| Participant Political Orientation x Uncorrected Accusation (vs. Neutral) Stimulus | 0.019 | 1.755 | 0.079 |
| Candidate Political Orientation x Uncorrected Accusation (vs. Neutral) Stimulus | 0.016 | 1.514 | 0.130 |
| **Affective Polarization Score x Uncorrected Accusation (vs. Neutral) Stimulus** | **-0.090** | **-7.978** | **< 0.0001** |
| **Participant Political Orientation x Candidate Political Orientation x Affective Polarization Score** | **0.030** | **2.147** | **0.032*** |
| Participant Political Orientation x Candidate Political Orientation x Corrected Accusation (vs. Neutral) Stimulus | -0.013 | -1.123 | 0.261 |

| | | | |
|---|---|---|---|
| Participant Political Orientation x Affective Polarization Score x Corrected Accusation (vs. Neutral) Stimulus | 0.021 | 1.827 | 0.068 ~ |
| Candidate Political Orientation x Affective Polarization Score x Corrected Accusation (vs. Neutral) Stimulus | -0.006 | -0.523 | 0.601 |
| Participant Political Orientation x Candidate Political Orientation x Uncorrected Accusation (vs. Neutral) Stimulus | -0.017 | -1.485 | 0.138 |
| Participant Political Orientation x Affective Polarization Score x Uncorrected Accusation (vs. Neutral) Stimulus | 0.015 | 1.301 | 0.193 |
| Candidate Political Orientation x Affective Polarization Score x Uncorrected Accusation (vs. Neutral) Stimulus | -0.010 | -0.857 | 0.392 |
| Participant Political Orientation x Candidate Political Orientation x Affective Polarization Score x Corrected Accusation (vs. Neutral) Stimulus | 0.001 | 0.098 | 0.922 |
| Participant Political Orientation x Candidate Political Orientation x Affective Polarization Score x Uncorrected Accusation (vs. Neutral) Stimulus | -0.021 | -1.800 | 0.072 ~ |

**Experiment 2 (n = 440)[1]**

| **Predictor** | **β** | **t** | **p** |
|---|---|---|---|
| **Initial Rating** | **0.140** | **15.224** | **< 0.0001*** ** |
| **Participant Political Orientation** | **-0.082** | **-3.501** | **0.0005*** ** |
| Candidate Political Orientation | -0.012 | -0.962 | 0.336 |
| Affective Polarization Score | 0.015 | 0.592 | 0.554 |
| **Corrected Accusation (vs. Neutral) Stimulus** | **-0.131** | **-15.252** | **< 0.0001*** ** |
| **Uncorrected Accusation (vs. Neutral) Stimulus** | **-0.510** | **-59.117** | **< 0.0001*** ** |
| Counterbalance1 | 0.022 | 0.966 | 0.335 |
| Counterbalance2 | 0.002 | 0.102 | 0.919 |
| **Participant Political Orientation x Candidate Political Orientation** | **0.090** | **6.777** | **< 0.0001*** ** |
| Participant Political Orientation x Affective Polarization Score | 0.028 | 1.104 | 0.27 |

[1] Note that contrary to our primary analysis, this mixed model analysis indicates an interaction between affective polarization and corrected accusation candidate status in Experiment 2 comparable to that observed in Experiment 1. However, interpretation of this effect is difficult due to the complexity of the model, i.e., interactions with candidate political orientation and participant political orientation. When those additional factors are excluded from the mixed effects model, the relationship between affective polarization and corrected accusation status is not significant, β = -.009, *t* = -0.877, p = .38. Further work will be necessary in the future to clarify the nature of this effect.

| | | | |
|---|---|---|---|
| Candidate Political Orientation x Affective Polarization Score | 0.007 | 0.548 | 0.583 |
| Participant Political Orientation x Corrected Accusation (vs. Neutral) Stimulus | -0.006 | -0.588 | 0.557 |
| **Candidate Political Orientation x Corrected Accusation (vs. Neutral) Stimulus** | **-0.021** | **-1.962** | **0.0497*** |
| **Affective Polarization Score x Corrected Accusation (vs. Neutral) Stimulus** | **-0.029** | **-2.639** | **0.008**** |
| Participant Political Orientation x Uncorrected Accusation (vs. Neutral) Stimulus | 0.015 | 1.383 | 0.167 |
| Candidate Political Orientation x Uncorrected Accusation (vs. Neutral) Stimulus | 0.020 | 1.935 | 0.053 ~ |
| **Affective Polarization Score x Uncorrected Accusation (vs. Neutral) Stimulus** | **-0.043** | **-3.841** | **0.0001***** |
| **Participant Political Orientation x Candidate Political Orientation x Affective Polarization Score** | 0.003 | 0.21 | 0.833 |
| **Participant Political Orientation x Candidate Political Orientation x Corrected Accusation (vs. Neutral) Stimulus** | **-0.025** | **-2.371** | **0.018*** |
| **Participant Political Orientation x Affective Polarization Score x Corrected Accusation (vs. Neutral) Stimulus** | **-0.034** | **-3.085** | **0.002**** |
| Candidate Political Orientation x Affective Polarization Score x Corrected Accusation (vs. Neutral) Stimulus | -0.003 | -0.275 | 0.783 |
| **Participant Political Orientation x Candidate Political Orientation x Uncorrected Accusation (vs. Neutral) Stimulus** | **-0.036** | **-3.341** | **0.0008***** |
| **Participant Political Orientation x Affective Polarization Score x Uncorrected Accusation (vs. Neutral) Stimulus** | **-0.025** | **-2.202** | **0.028*** |
| Candidate Political Orientation x Affective Polarization Score x Uncorrected Accusation (vs. Neutral) Stimulus | 0.006 | 0.544 | 0.587 |
| Participant Political Orientation x Candidate Political Orientation x Affective Polarization Score x Corrected Accusation (vs. Neutral) Stimulus | -0.001 | -0.102 | 0.919 |
| Participant Political Orientation x Candidate Political Orientation x Affective Polarization Score x Uncorrected Accusation (vs. Neutral) Stimulus | -0.007 | -0.575 | 0.565 |

Supplemental Table S10. Mixed effects models examining effects of perceived candidate ideology, participant ideology, and stimulus condition on immediate ratings.

Model specification: Immediate Post-Story Rating ~ Initial Pre-Story Rating + Participant Ideology * Candidate Ideology * Corrected Accusation + Participant Ideology * Candidate_Ideology * Uncorrected Accusation + Counterbalance1 + Counterbalance2 + (1|Subject)

**Experiment 1 (n = 454)**

| **Predictor** | **β** | **t** | **p** |
|---|---|---|---|
| **Initial Rating** | **0.170** | **18.760** | **< 0.0001*** ** |
| **Participant Political Orientation** | **-0.062** | **-2.585** | **0.010** ** |
| Candidate Political Orientation | -0.022 | -1.825 | 0.068 ~ |
| **Corrected Accusation (vs. Neutral) Stimulus** | **-0.113** | **-13.784** | **< 0.0001*** ** |
| **Uncorrected Accusation (vs. Neutral) Stimulus** | **-0.442** | **-53.989** | **< 0.0001*** ** |
| Counterbalance1 | 0.007 | 0.273 | 0.784 |
| Counterbalance2 | 0.012 | 0.482 | 0.630 |
| **Participant Political Orientation x Candidate Political Orientation** | **0.059** | **4.749** | **< 0.0001*** ** |
| Participant Political Orientation x Corrected Accusation (vs. Neutral) Stimulus | -0.002 | -0.199 | 0.843 |
| Candidate Political Orientation x Corrected Accusation (vs. Neutral) Stimulus | -0.009 | -0.944 | 0.345 |
| **Participant Political Orientation x Uncorrected Accusation (vs. Neutral) Stimulus** | **0.039** | **3.925** | **< 0.0001*** ** |
| **Candidate Political Orientation x Uncorrected Accusation (vs. Neutral) Stimulus** | **0.027** | **2.668** | **0.008** ** |
| Participant Political Orientation x Candidate Political Orientation x Corrected Accusation (vs. Neutral) Stimulus | -0.012 | -1.196 | 0.232 |
| Participant Political Orientation x Candidate Political Orientation x Uncorrected Accusation (vs. Neutral) Stimulus | -0.018 | -1.760 | 0.078 ~ |

**Experiment 2 (n = 499)**

| **Predictor** | **β** | **t** | **p** |
|---|---|---|---|
| **Initial Rating** | **0.151** | **17.567** | **< 0.0001*** ** |
| **Participant Political Orientation** | **-0.062** | **-2.952** | **0.003** ** |
| Candidate Political Orientation | -0.012 | -0.962 | 0.336 |
| **Corrected Accusation (vs. Neutral) Stimulus** | **-0.119** | **-14.910** | **< 0.0001*** ** |
| **Uncorrected Accusation (vs. Neutral) Stimulus** | **-0.497** | **-61.951** | **< 0.0001*** ** |
| Counterbalance1 | 0.015 | 0.667 | 0.505 |
| Counterbalance2 | -0.005 | -0.212 | 0.833 |

| | | | |
|---|---|---|---|
| **Participant Political Orientation x Candidate Political Orientation** | **0.082** | **6.736** | **< 0.0001*** ** |
| Participant Political Orientation x Corrected Accusation (vs. Neutral) Stimulus | -0.011 | -1.126 | 0.260 |
| **Candidate Political Orientation x Corrected Accusation (vs. Neutral) Stimulus** | **-0.024** | **-2.466** | **0.014*** |
| Participant Political Orientation x Uncorrected Accusation (vs. Neutral) Stimulus | 0.015 | 1.582 | 0.114 |
| Candidate Political Orientation x Uncorrected Accusation (vs. Neutral) Stimulus | 0.018 | 1.806 | 0.071 ~ |
| **Participant Political Orientation x Candidate Political Orientation x Corrected Accusation (vs. Neutral) Stimulus** | **-0.022** | **-2.217** | **0.027*** |
| **Participant Political Orientation x Candidate Political Orientation x Uncorrected Accusation (vs. Neutral) Stimulus** | **-0.037** | **-3.778** | **0.0002**** |

Supplemental Appendix. Digital literacy measure. Correct answers are bolded.

1. How are decisions about what stories to show people on Facebook made?[2]
    a. At random
    b. By editors and journalists that work for news outlets
    c. By editors and journalists that work for Facebook
    **d. By computer analysis of what stories might interest you**

2. On Twitter, which kind of tweets will you NOT see in your feed?
    a. Retweets shared by people you follow
    b. Sponsored advertisements
    c. Tweets replied to by people that you follow
    **d. Tweets from accounts that follow you, but you don't follow back**

3. If you do not want to see a certain category of content on TikTok (not just from one account), what can you do to make similar content less likely to show up in your feed?
    a. Go to the account that posted it and click on the bell symbol at the top right corner
    b. Click on the arrow menu next to the post and select "report post"
    **c. Click on the arrow menu next to the post and select "not interested"**
    d. Go to the account that posted it, click on the three-dot menu at the top, and click "block"

4. What is phishing?
    **a. Sending fraudulent messages pretending to be from reputable companies in order to get individuals to reveal personal information**
    b. Luring someone into a relationship online through a fictional online persona
    c. When cyber-criminals hack into a computer network to extract sensitive information
    d. Embedding ads into website through layering them on top of each other so that they are not visible, but you may accidentally click on them

5. When you block someone on a social media site, what effect does it NOT have?
    a. They can't see your posts on the platform
    b. You can't see their posts on the platform
    **c. They are banned from the platform for a period of time**
    d. They can't see your engagement with accounts that you both follow

6. What is the term for when someone references you in a post and it sends a notification to you and/or your followers?
    **a. Tagging**
    b. Reference
    c. Friending
    d. Direct message

[2] Question taken from Sirlin et al. (2021)