The role of conflict and alignment in shaping responses to polarizing factchecks

Word count:

Fletcher Scott[1], Damiano Spina[2], and Lauren Saling[1]

[1]School of Health and Biomedical Sciences, RMIT University

[2]School of Computing Technologies, RMIT University

Author Note

There are no conflicts of interest to disclose.

This study does not reproduce any material requiring permission.

The data, analysis code, and supplementary materials for this project are openly available at the Open Science Framework (OSF). You can access these materials here: https://osf.io/waehy/?view_only=0662484aaaf5406ca9a892ec2543fc51

Funding

This research was funded by the School of Health and Biomedical Sciences, RMIT University and the ARC Centre of Excellence for Automated Decision-Making and Society.

Author Contributions

Fletcher Scott; conceptualization, project administration, data curation, formal analysis, investigation, methodology, software, writing – original draft.

Damiano Spina; writing – supervision, review and editing

Lauren Saling; conceptualization, project administration, supervision, writing – review and editing

Competing Interests

There are no conflicts of interest to disclose.

Ethics Approval and consent to participate

This study was approved by the STEM College Human Ethics Advisory Network (CHEAN), RMIT University, under approval number 26411. All procedures were conducted in accordance with the guidelines and regulations set forth by this board.

Permission to reproduce materials

This study does not reproduce any material requiring permission.

Preregistration statement

This study was not preregistered.

Abstract

People often continue to endorse claims after fact-checks (the continued influence effect, CIE). Dominant accounts attribute CIE to a failure to detect conflict between prior beliefs and corrective evidence, or because, once they do notice, they respond in a reactive partisan-consistent way. In a 2×2 (Congruence × Truth) online experiment with N=200 Australian adults, we measured participant evaluations at two levels: rapid associations (Brief Implicit Association Test; BIAT) and normative judgments (self-report). We predicted that larger implicit–explicit discrepancies (IED) would magnify partisan-congruent processing if an identity-protective form of dissonance dominated. Contrary to prediction, while perceptions of statements were clearly shaped by partisanship and fact-checking robustly predicted belief updating, IED did not reliably moderate the degree of belief updating. Instead, higher overall source liking, indicated by alignment across measures, broadly dampened corrective shifts irrespective of congruence, suggesting that coherent intuitions about both preferred and nonpreferred sources set the weight of correction, not conflict between them. This pattern is consistent with conflict being registered at the statement–evidence level while source-level discrepancy remained inert. We propose that to increase item-level accuracy, corrections should generally suppress identity coactivation and keep sources peripheral so that judgments anchor themselves in statement-evidence coherence.

Keywords: political partisanship, fact-checking, source credibility, Implicit Association Test, continued influence effect

The role of conflict and alignment in shaping responses to polarizing factchecks

Introduction

Partisan deployment of misinformation undermines democratic functioning because it leads to beliefs that appear to reflect genuine conviction rather than calculated deception (Zhang & Rand, 2023). Once framed as relevant to political identity, information processing becomes partly driven by morally defensive responses (Tappin & McKay, 2019; Kahan, 2017; Rathje et al., 2021), linked with a biased form of reasoning about facts (Drummond & Fischhoff, 2017; Pretus et al., 2023). This has documented effects on key choices regarding U.S. elections (Guess et al., 2020) and COVID-19 (Pennycook et al., 2020). In such contexts, people often rely on false but partisan-consistent information when deciding how to vote or what to share, amplifying polarization and collective exposure (Chen et al., 2021; Osmundsen et al., 2021).

Fact-checking improves item-level accuracy but rarely reverses shifts in both the targeted belief and the wider attitudes it underpins (Swire et al., 2017; Thorson, 2016), known as the continued influence effect (CIE; for reviews, see Ecker et al., 2022; Lewandowsky et al., 2012). The prevailing view is that consistent fact-checking exposes conflicts between prior beliefs and new evidence, acting as a signal to engage analytic scrutiny (Bottoms et al., 2010; Bronstein & Pennycook, 2019). This process is typically explained through dual-process theory, which holds that people default to fast, intuitive (“Type 1”) judgments, while slower, analytic (“Type 2”) reasoning is recruited only when a conflict-monitoring system detects competing intuitions (Pennycook, 2023).

Whether this activates depends in part on a general tendency to conserve mental effort, making those less willing to invest cognitive resources more likely to accept their intuitions (Pennycook & Rand, 2019, 2021; Ecker & Ang, 2019; Pretus et al., 2023). Inspired by the finding that beliefs of strong partisans are particularly difficult to correct (Thorson, 2016), concerns arise regarding whether fact-checks may indeed prompt people to notice a clash between prior beliefs and new evidence, and yet resolve those conflicts by relying on partisanship as a default solution. This aligns with work suggesting that corrections leave an explanatory gap that sustains conflict with the prior narrative, making them effective only when they offer a plausible alternative (Lewandowsky, 2005; Walter & Tukachinsky, 2020).

We argue that interventions should be evaluated on their capacity to resolve conflicting intuitions rather than merely reveal them (van Harreveld et al., 2015). To do so, conflict is operationalized as the implicit–explicit discrepancy (IED) between evaluations of the same partisan sources (Briñol et al., 2006; Petty et al., 2006). Divergence is expected because implicit and explicit measures index different representational formats. Implicit evaluations reflect associative activation shaped by fluency, whereas explicit reports reflect propositional endorsements shaped by goals and norms (Gawronski & Bodenhausen, 2006; Payne et al., 2008; Strack & Deutsch, 2004; Wilson, Lindsey, & Schooler, 2000). Under identity concerns, explicit judgments can remain stable while implicit associations shift with evidence-driven fluency signals, or the reverse when motivated reasoning constrains endorsement (Fazio, 1990; Dechêne et al., 2010; Unkelbach & Greifeneder, 2013). Our central test is whether higher IED at the moment of correction predicts reduced movement toward corrective evidence, resulting from stronger maintenance of partisanconsistent judgments.

Conflict detection and resolution

Current dual-process models propose that upon observing a news headline, Type 1 processing generates initial responses that are perceived as ready-made answers (Evans & Stanovich, 2013). These intuitions arise through quick pattern recognition, where previously encountered cues re-activate familiar representations (Pennycook et al., 2015). However, the individual can intervene in these shortcuts by engaging in Type 2 processing (Meyer & Frederick, 2023). Dominant models suggest that this is achieved by maintaining multiple competing concepts and adjudicating among them (Chaiken & Trope, 1999; De Neys, 2025), a process supported by working memory and the inhibition of pre-potent responses (McIlhiney et al., 2023). Empirically, this appears to operate by sustained elaboration of a set of chosen intuitions rather than by generating new, alternative outputs (Beauvais et al., 2025). As a result, any conflict between Type 1 impressions and the diagnostic information justifying those impressions becomes salient, making revision more likely (Petty & Cacioppo, 1986; Pennycook, 2017).

Across misinformation studies, analytic (Type 2) reasoning predicts better truth discernment, aligning with the view that analytic interventions on initial intuitions improves accuracy (Bronstein et al., 2019; Pehlivanoglu et al., 2021; Pennycook & Rand, 2019, 2020). Consistent with dual-process accounts, revision is most likely when conflicting representations are coactivated. Corrections that re-cue the original claim improve accuracy chiefly when people later retrieve both the misinformation and its correction (Walter & Tukachinsky, 2020). Without such recollection, retrieval is instead linked to reduced accuracy (Kemp et al., 2022, 2024). Analytic thinkers indeed show that such conflicts affect reasoning, exhibiting longer response times relative to less analytic individuals and relative to their own responses to non-contradictory information (Pennycook et al., 2014; De Neys, 2012). Building on this, interventions that surface conflict or extend deliberation time have been shown to improve truth discernment, as longer decision windows or opportunities to revise intuitions hypothetically allow competing representations to be maintained and arbitrated (Bago et al., 2020; Sultan et al., 2022; De Neys, 2012, 2014, 2023).

A key concern, however, is that conflict detection may occur without leading to correction. Evidence shows people can sense conflict and yet still rely on their initial intuitive decision (Denes-Raj & Epstein, 1994; De Neys, 2006; De Neys & Glumicić, 2008; Pennycook et al., 2014; Frey et al., 2018), resulting in dissociations under identity or effort constraints (De Neys & Glumicić, 2008; De Neys, 2012; Pennycook et al., 2014). Classic base-rate problems illustrate this disconnect. These tasks require integrating prior category probabilities (base rates) with case-specific evidence. Researchers find that when a stereotypical description of personal traits or categories contradicts a statistical base rate, participants slow down and report lower confidence during decision-making. This pattern is taken as evidence that detection has occurred and Type 2 processing has been engaged. However, this detection notably often fails to translate into correction, as many participants still provide the intuitive, normatively incorrect response (De Neys & Glumicic, 2008; De Neys, 2012; Pennycook et al., 2014). Researchers interpret this as evidence of a monitoring–control gap, or a stage in which, under the uncertainty created by detected conflict, there is a decision made about which cues to treat as diagnostic for resolving that conflict (Gratton et al., 1992; Botvinick et al., 2001; Kerns et al., 2004; Egner, 2007), presented in figure 1.

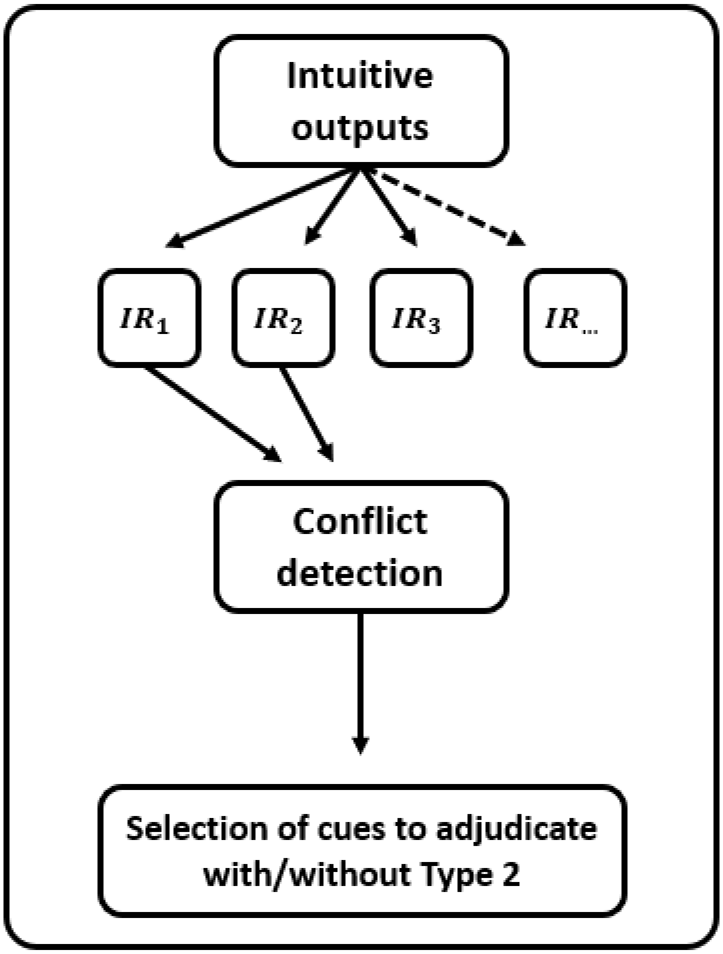

Fig. 1: Type-1 conflict-detection model of credibility inference. Fast Type-1 processes generate initial intuitions about a statement and source; a domain-general conflict monitor flags inconsistencies among them. Type-2 processing expands the set of candidate intuitions, allowing the same monitor to re-check coherence and promote updates toward more epistemically accurate beliefs. This figure is adapted from the Three-stage dual-process model of analytic engagement (Pennycook, 2023).

The particular concern for political contexts is that conflict detection appears to take place as part of a Type 1 system, before Type 2 control can intervene. Notably, such markers of conflict (e.g., lower confidence and longer reaction times) persist even under tight deadlines and high cognitive load (Bago et al., 2017; Bago & De Neys, 2020; Johnson et al., 2016), conditions that diminish control over prepotent responses. Neurocognitive findings indicate that conflict is most notably registered through pathways dedicated to processing identity-relevant information, particularly when the detected conflict is in response to risk-related and moral-emotional language (Leong et al., 2020), and that these arrive as early signatures in higher-order and sensory networks for partisans (Katabi et al., 2023). Consistent with this view, conflict is typically felt as aversive affect that rapidly reallocates attention and shifts cue weighting, whereas processing fluency carries a positive affect that functions as a validity cue, boosting associative accessibility and thereby biasing endorsement (Stump et al., 2022; Dignath et al., 2020; Martel et al., 2020; Dechêne et al., 2010; Unkelbach & Greifeneder, 2013; Fazio et al., 2015; Pennycook et al., 2018; Brashier & Marsh, 2020). These early signals may be useful to explain downstream partisan behaviours, who subsequently adjudicate identity-relevant cues and show stronger CIE (Baum & Abdel Rahman, 2021; Bullock, 2009; Druckman & McGrath, 2019; Hill, 2017; McLaughlin et al., 2020; Petersen et al., 2015).

Although a precise characterization of this stage is lacking, conflict detection in Type 1 processing may bias subsequent processing to the conflict being detected, and if one exists regarding their identity, partisans more readily attend to ways that could resolve these identity-relevant intuitions. This does not mean that partisans think under a different mechanism, or fail to understand the data, but represent the overall problem differently and thus work from different intuitions (see Figure 2). This is a departure from classic motivated-reasoning accounts in which partisans use Type 2 processing to defend their identities (Kahan et al., 2017). In this view, a domain-general conflict monitor is especially sensitive to identity-relevant clashes, because partisan contexts readily generate strong, identity-consistent intuitions (Gawronski et al., 2023).

For example, Van Boven et al. (2019) shows that when the same four statistics are presented in a contingency table, partisans disagree about which one “matters,” a choice that heightens susceptibility to misinformation. A similar experiment replicated the effect, with political orientation biasing gaze toward different graph regions, consistent with early selection setting what is diagnostic before Type 2 (Luo & Zhao, 2019; Yu & Opfer, 2024).

Yet, such studies have not tested whether these findings were the result of early conflict signals. Finding an answer is important because it clarifies whether and how early conflict signals drive revision or entrenchment, and thus whether interventions should surface or avoid specific conflicts as a method of reshaping the intuitions being adjudicated. In line with this, Wischnewski & Krämer (2021) show that momentary affect, rather than identity by itself, predicted utilization of identity in truth judgements, particularly states of anxiety. This potentially explains why clear, specific corrections reduced reliance on misinformation to a similar extent among Democrats and Republicans (Ecker et al., 2022), yet when a state of hostility is induced, identity-threat cues activate a group-oriented form of conflict, so partisans discount counter-attitudinal evidence and diverge in belief accuracy even under identical inputs (Kim, 2025; Su, 2022).

In this study, we investigated the role of conflict monitoring in the CIE to understand how people draw inferences from these signals when revising their beliefs. Despite robust evidence that people sense conflicts between prior beliefs and new evidence even under time pressure, most work does not isolate what happens after conflict is detected. In polarized contexts, identity cues can be privileged, but prior studies rarely (a) separate detection from resolution within persons, (b) test arbitration at the statement and source levels, or (c) use a process-sensitive index of person-specific conflict that can forecast the direction of updating. As a result, CIE is often attributed to detection failures.

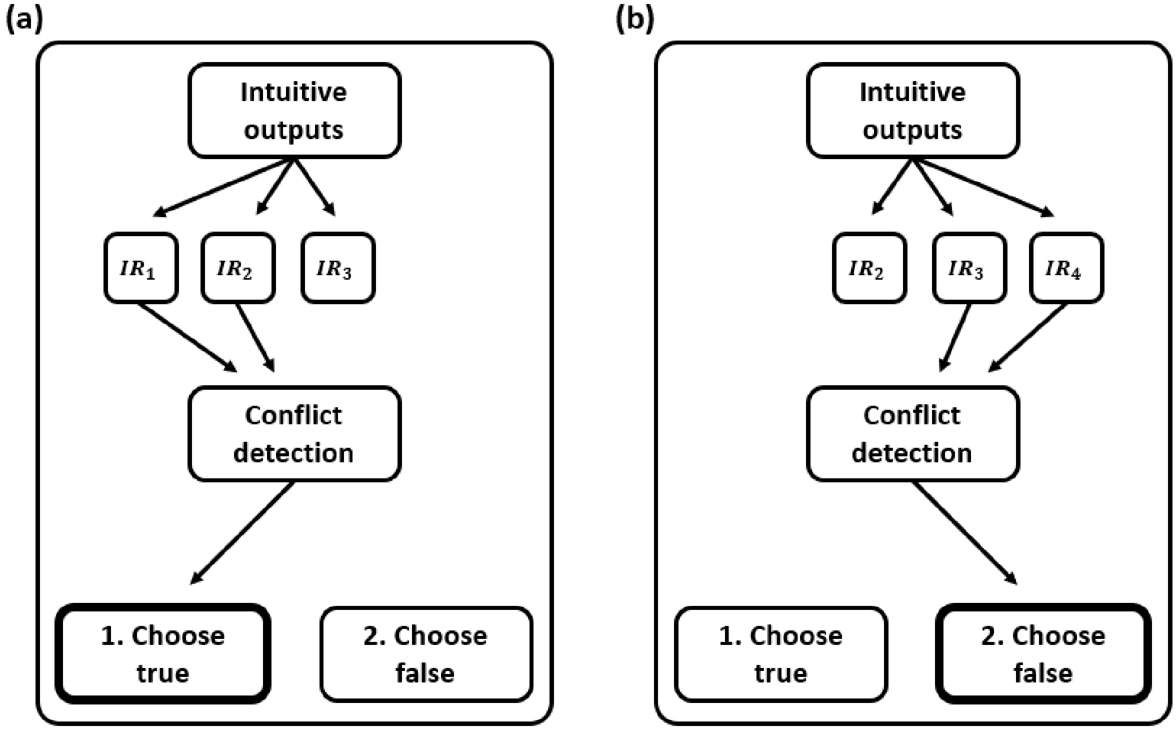

Fig.2: Pathways to credibility choice from Type-1 intuition sets. (a–b) A common conflict-detection mechanism evaluates the coherence of initial intuitions (IR_i) generated by Type-1 processing. Panels differ only in the structure of IR_i. In (a), partisan-congruent prior exposure and information consumption skew the available intuition set toward identity-relevant cues, so the same monitor resolves to “true.” In (b), a different or more mixed intuition set yields “false.” Type-2 deliberation is not modeled here; if engaged, it would expand the set of IR_i evaluated by the same detector, biased toward those initially activated. IR_i denotes intuition i.

The present study

The primary aim of this study is to investigate whether the gap between an individual’s intuited associative and normative reactions to the correction of partisan sources (IED) predicts how much and in which direction they change their belief after a fact-check. The measured outcome was each person's change in confidence after the fact-check for every statement, computed within a 2×2 design crossing Congruence (statement from one's own side vs the other side) and Truth (fact-check says true vs false). Positive change means increased endorsement after the correction. Given veracity sensitivity, we expect increases for true items and decreases for false items on average.

We measured source preferences in two ways—explicit (self-report) and implicit (BIAT-style). We summarize their gap as an explicit–implicit discrepancy (IED), standardized so that larger absolute values indicate a bigger mismatch, and positive values mean automatic evaluations are more favorable than stated views. We model explicit and implicit preferences jointly using response-surface terms that separate (a) overall liking of the source from (b) mismatch between explicit and implicit views. This avoids artifacts of raw difference scores while letting us ask whether mismatch, rather than liking, relates to updating.

The study hypothesized that people would shift toward the fact-check's verdict, but the size of that shift would be impacted by their partisanship. When the source is congruent with the participant's political affiliation, a larger conflict in representations should activate and thus magnify updates that favor that position. In comparison, when the correction harms the ingroup, mismatch should reduce the impact of corrections. We expect this pattern to grow with partisanship strength and to remain after accounting for overall liking.

Method

Participants

Data was collected online between the 31 August and 3 September 2023. Participants were recruited through Prolific, an online platform. They were compensated £3 for their participation, with a median completion time of 16 minutes (£11.25/hour). Consent was implied through the submission of responses. Eligibility required Australian residency, being 18 years or older, and use of a desktop or laptop device. Mobile/tablet participation was blocked to ensure IAT timing reliability.

An apriori power analysis using G*Power Version 3 for a mixed model ANOVA, alpha .05, power .8, revealed a minimum required sample size of 112. Following recommendations for using the Brief Implicit Association Test (BIAT) (Greenwald et al., 2022; Nosek et al., 2014) and anticipating potential data attrition, an initial sample of 225 Australian residents was recruited. Data were excluded if participants failed the attention check (1 item), the political-knowledge screen (2 items; both required for inclusion), or completed <50% of the study. Political knowledge questions were included to ensure that participants were actively engaged with Australian politics.

Of 225 participants, 25 were excluded per preregistered criteria: failed attention check (n = 15), failed political-knowledge question (n = 10), incomplete (<50%; n = 18). Reasons were not mutually exclusive. The final analysed sample was N = 200. The final sample comprised 98 males (49.0%), 97 females (48.5%), and 5 non-binary or gender diverse participants (2.5%). Ages ranged from 18 to 77 (M = 36.52, SD = 12.36; median = 34). Participants skewed leftleaning: political orientation calculation yielded 147 left (73.5%) and 53 right (26.5%).

Ideological strength was moderate on average (M = 0.41, SD = 0.21). See Table 1 for sample characteristics.

Table 1

Sample characteristics

Characteristic	n (%) or M ± SD [Median (IQR where reported), Range]
Age (years)	36.52 ± 12.36 [34, 18–77]
Gender	
Male	98 (49.0%)
Female	97 (48.5%)
Non binary / gender diverse	5 (2.5%)
Political orientation (5-item index; sum −10 to +10)	4.01 ± 3.94 [4 (1–8), −4–10]
Right (≤ −3)	10 (5.0%)
Centre (−2 to +2)	57 (28.5%)
Left (≥ +3)	133 (66.5%)
Self-identified Party Affiliation (0–100)	65.11 ± 21.77 [65 (50–81), 0–100]
Far Right (0-16)	4 (2.0%)
Right (17–33)	8 (4.0%)
Centre-Right (34–44)	21 (10.5%)
Centre (45–55)	43 (21.5%)
Centre-Left (56–66)	27 (13.5%)
Left (67–83)	52 (26.0%)

Far Left (84–100)	45 (22.5%)
Education	
Did not complete high school	1 (0.5%)
Year 12 or equivalent	38 (19.0%)
Diploma or equivalent	26 (13.0%)
Bachelor’s degree	85 (42.5%)
Post-graduate degree	50 (25.0%)
Completion time (minutes)	19.52 ± 12.29 [16.63 (13.82–21.25), 7.12–117.38]

Note: Political-orientation index: higher scores = more left-leaning (−10 = far right, +10 = far left). Analyses used a 6-item composite (5-item index + self-placement) after rescaling to a common metric; see Measures.

Measures

Political Orientation (composite index)

All items were recoded so that higher values indicate more left-leaning in orientation. The five attitude items (five-point Likert) were mapped to a common metric of −2 to +2 (Strongly Disagree = −2 … Strongly Agree = +2), with two items reverse-coded based on their directional wording. The 0–100 self-placement was inverted and linearly rescaled to −2 to +2. We averaged the six recoded items and rescaled the mean to −10 to +10 by multiplying by 5 (−10 = far right, +10 = far left). Internal consistency (six-item composite) was α = .85. For interpretability, Table 1 reports descriptives for the five-item index (−10…+10) and the self-placement (0–100) separately.

Political Statements, Confidence Ratings, And Fact Checks

Stimuli comprised eight political statements, four per politician (two true, two false), attributed to Anthony Albanese (Labor) and Peter Dutton (Liberal), all drawn from items previously reviewed by RMIT ABC Fact Check. Presentation order was randomized. After each statement, participants reported their confidence that the statement is true on a 0–100 slider (higher = more confident true).

After the initial rating phase, each statement was followed by a standardized fact-check comprising: (i) the original statement, (ii) a True/False verdict, (iii) 2–3 sentences of justifying evidence, and (iv) an onscreen reminder of the participant's prior confidence rating. Participants then re-rated their confidence for each statement on the same 0–100 scale. The primary outcome was ΔConfidence = Post − Pre at the trial level.

Explicit Source Evaluations (E)

Following the fact-checking phase, participants rated each politician on two items adapted from Aird et al. (2018): overall credibility and perceived factual accuracy/truthfulness.

For each politician separately, item scores were z-scored across participants and averaged to form explicit composites EAlb and EDut (higher = more favorable explicit evaluation of that politician as an information source).

Implicit Preference for Sources (I)

Additionally, following the fact-checking stage and having rated sources with an explicit measure, we used a BIAT to measure a different type of intuition generated about political sources and credibility. Following Sriram & Greenwald (2009) and Nosek et al. (2014), faster/more accurate pairings indicate stronger associations. An adapted Project Implicit script (Bar-Anan, 2020) ran in Qualtrics via Minno.js (Zlotnick et al., 2015). Participants sorted stimuli from two political sources (Albanese, Dutton; see Table 2) using the I/E keys. Emulating Greenwald et al., (1998), attributes were evaluative states of credibility rather than affective. This was to equip the BIAT with trait dimensions that associate most strongly with credibility at the explicit level (Appelman & Sundar, 2016). Five target images were selected from large-scale Australian media corporations. The split-half reliability was found to be r = 0.78.

Table 2

BIAT procedure

Note: Blocks alternated focal pairings (e.g., Albanese or Honest vs. Dutton or Lying) with the complementary pairing; block order and key assignment (I/E) were counterbalanced. Each block began with 4 practice trials followed by 16 test trials; only test trials were analyzed (total test trials = 64). Attribute word lists (credibility vs. non-credibility) and target images are provided in the appendix/OSF.

Latencies <300 ms or >10,000 ms were excluded. Incorrect responses were replaced with the corrected latency plus 600 ms, and D-scores were computed following Sriram and Greenwald (2009). Positive DBIAT values indicate faster responding on blocks pairing Albanese + Honest (relative to Dutton + Honest), indexing a stronger implicit association between Albanese and credibility. To align the implicit metric with the speaker on each trial, we mapped the participant-level DBIAT to a trial-level score I by reversing the sign as needed: for Albanese trials I = + DBIAT; for Dutton trials I = − DBIAT. Thus, higher I always reflects a more favorable implicit evaluation of the trial's speaker.

Analysis Coding and Moderators

Explicit (E) and implicit (I) evaluations were z-scored across participants within measure and rotated to elevation and discrepancy axes. We used L (overall liking) and D^2 (magnitude of explicit–implicit disagreement) as focal predictors. For descriptive purposes only, we also computed IED = z(I) − z(E) and |IED| per participant × source; neither term entered confirmatory models to avoid difference-score artifacts.

Results

IED

We indexed IED at the participant × source level, capturing the absolute magnitude of discrepancy regardless of direction. Across all cases, the average absolute discrepancy was moderate in size (M = 0.87, SD = 0.67; Median = 0.73). In total, 63.75% of cases exceeded a 0.50 SD benchmark and 36.25% exceeded a 1.00 SD benchmark. Cases in which explicit and implicit evaluations had opposite signs ($E \cdot I < 0$) occurred in 36.25% of cells.

As an index of E–I alignment, Pearson correlations between explicit and implicit evaluations were positive within both sources, indicating modest correspondence: Alb r = 0.468, Dut r = 0.334; overall alignment was r = 0.403.

Identity-consistent processing

The observed SD of Δ\DeltaΔ was 40.13. We prespecified a smallest effect size of interest (SESOI) of ±0.05 SD, i.e., ±2.006 points on the 0–100 scale, and used TOST to evaluate equivalence for the discrepancy interactions.

We fit linear mixed-effects models (SPSS) with fixed effects of Congruence, Truth, L (z), D^2, and P (z), and their interactions. Our confirmatory tests were H1, namely Congruence × D^2, or that more implicit-explicit disagreement predicts more partisan-aligned updating and H1m, that the pattern of Congruence × D^2 × P is stronger with higher partisanship strength. We also probed Congruence × L to test whether overall liking (where E and I move together) could explain partisan-aligned updating. A random intercept for statement was retained. A participant random intercept failed to converge and had near-zero variance, but the statement-only model converged and improved fit of −2REML LL from 15087.94 to 14873.29.

As expected, fact-checks labelled True increased confidence and False decreased it, $F \approx 44.92$, $p < .001$, with a significant Congruence × Truth interaction, $F \approx 7.27$, $p = .007$. The interaction followed the expected pattern (True–congruent > True–incongruent; False–incongruent < False–congruent).

There was no evidence that discrepancy predicted partisan-aligned updating (Congruence × D^2: $F\approx0.81$, $p=.369$; $b=-1.446$, $SE=1.609$, 95% CI [−4.603, 1.710]) nor that this effect was moderated by partisanship strength (Congruence × D^2 × P: $F\approx0.51$, $p=.600$). Using ±2.006 as the equivalence bounds (±0.05 SD), TOSTs for Congruence × D^2 and Congruence × D^2 × P did not pass both one-sided tests, so the discrepancy effects were non-significant but not demonstrably negligible. Instead, we observed a main effect of higher overall liking, which predicted smaller belief updates on average, $F(1,\approx1582)=7.23$, $p=.007$. The simple slope at the reference congruence level was $b = -1.59$ ($SE = 0.94$), 95% CI [−3.44, 0.26], indicating ~1–2 points less updating per SD increase in L on the 0–100 scale. Congruence × L was not significant, $F(1,\approx1580)=0.19$, $p=.665$, meaning that liking did not differentially affect congruent vs. incongruent trials and cannot explain partisan-aligned processing. These results are reported from the converged model with a random intercept for statement. The SD of updating was 40.13 on the 0–100 scale; our SESOI of ±0.05 SD corresponds to ±2.006 points.

Discussion

Our primary aim was to test whether the discrepancy between participants' explicit and implicit evaluations of partisan sources (IED) predicts the magnitude and direction of belief revision after fact-checks, over and above overall source liking, and whether any such effect scales with partisanship strength. This hypothesis was not supported. The present study found that corrections reliably shifted itemlevel beliefs toward veracity. Confidence increased for statements labelled ‘true’ and decreased for those labelled ‘false,’ though its magnitude was reduced for partisan-congruent statements. Our results therefore replicate the CIE, with political congruence moderating that effect. Although IED was present following fact-checks, it did not moderate the size or direction of updating. Instead, independent of congruence, higher overall source liking dampened corrective shifts, indicating that identity-linked signals did influence arbitration but through alignment rather than through conflict. Put differently, coherent intuitions about both preferred and non-preferred sources set the weight of correction, not conflict across intuitions. Taken together, the pattern suggests that arbitration was governed primarily at the statement–evidence level, while source-level inconsistency remained behaviourally silent and source-level alignment (liking/fluency) lowered the gain assigned to corrective evidence. In practice, higher liking likely indexes stronger pre-existing priors, a positive association that, on average, reduces openness to revision for both ‘true’ and ‘false’ verdicts.

Where was conflict registered?

Participants adjusted confidence toward fact-check veracity, indicating that new evidence was processed. Yet the conflict signal that entered arbitration appears to have been local to the statement (statement–evidence coherence) rather than about the source (source honesty/reputation). Although implicit–explicit discrepancy (IED) was present, it did not moderate updating, whereas higher overall source liking dampened corrective shifts irrespective of congruence. Thus, identity- linked cues did influence arbitration, but via elevation/alignment (a shared direction of explicit and implicit intuitions) rather than via conflict between them. Put differently, coherent intuitions about both preferred and non-preferred sources set the weight placed on corrective evidence while mismatch did not.

We interpret these findings through a conflict-registration account. Multiple conflicts can be available at once (statement–evidence, source–reputation, evidence–identity), but only registered conflicts shape arbitration on the corresponding dimension. Registration depends on coactivation and link strength among the cues defining that dimension. In the present design, the monitor likely registered statement–evidence conflict while source-level conflict did not achieve sufficient coactivation to affect triallevel changes in confidence, leaving IED inert as a moderator. Meanwhile, alignment/elevation (liking/fluency) acted as a gain control on evidence, so that higher elevation meant that a single corrective message carried less weight. Crucially, because liking did not interact with congruence, it cannot account for the partisan tilt observed. Congruence thus likely to drew on a range of cues available to resolve the statementlevel contradiction.

A Congruence × Truth interaction indicated a partisan tilt in updating. A Bayesian perspective explains this without invoking reasoning failure, where subgroups may hold different priors over source reliability and claim types, so the same likelihood (factcheck) can yield different posterior movements. In everyday prediction, people recruit domain-appropriate priors that mirror realworld statistics, yielding nearoptimal predictions that nevertheless differ across domains (e.g., Griffiths & Tenenbaum, 2006). With distinct priors but the same likelihood, formal analyses show that polarization can emerge under normative Bayesian updating (Jern, Chang, & Kemp, 2009, 2014), so divergences do not, by themselves, imply a lazy form of updating. Empirically, the prior-attitude effect (Taber & Lodge, 2006), often strongest among individuals with strong priors (Kahan et al., 2017), fit the idea that differently shaped prior exposures dampen updating unless evidence is unusually diagnostic. Viewed this way, the partisan tilt we observe reflects mismatched prior distributions over claims.

Several boundary conditions plausibly account for the null moderation without denying person-level inconsistency having occurred. For source conflict to influence arbitration, the discrepancy may need to be both large and made salient. In our materials, contradictions were likely low-threat and could be resolved locally at the statement level without considering and resolving to source credibility. In addition, collapsing discrepancy to a single magnitude (|E−I|) conflates two regimes that may have opposite implications for registration—E > I (declared positivity with weak associative pull) versus I > E (associative positivity with restrained endorsement). Pooling them may have disguised direction-specific effects. A further constraint is a level-of-analysis mismatch. IED was measured globally (person × source), whereas updating was expressed locally (trial × statement), making it plausible that global mismatch governs slower reputation drift rather than immediate responses to a particular correction. Finally, the presentation of task elements likely reduced the simultaneous coactivation of source and statement-level representations required for source conflict to register. Without such coactivation, source inconsistency can remain latent even when it exists.

We propose that progress on CIE requires modeling the locus of conflict registration. To that end, we introduce Relational Conflict Distance (RCD), defined as the graded referential distance between a source-linked intuition and a statement-linked intuition. When RCD is low, coactivation is easy and source conflict is more likely to register alongside statement-level conflict. When RCD is high, the system can resolve the statement locally without ever registering a source-level inconsistency, meaning that a potentially extant IED remains behaviorally silent. This view aligns with work showing that perceived validity generalizes from exact repeats to paraphrases and variants, and can even reemerge for delayed contradictions, consistent with activation traveling along learned links (Silva, Garcia-Marques, & Reber, 2017; Garcia-Marques, Silva, Reber, & Unkelbach, 2015; Unkelbach & Rom, 2017). Similarly, partisans find it easier to accept a correction regarding a sole member of their party over one regarding their party as a whole (Puryear et al., 2024).

Taken together, these findings shift the objective from simply surfacing conflict to managing which conflict is considered relevant and how much weight corrective evidence receives. When the goal is item-level accuracy, interventions should keep identity cues latent so that arbitration remains at the statement–evidence level: present the evidence first, minimise or delay party labels and logos, and avoid repeated source mentions across paraphrases that could pull source cues into arbitration. When the goal is reputation updating, the strategy reverses and fact-checkers should deliberately coactivate source and statement so that source-level inconsistency is registered. This could potentially occur by colocating source tags with claims, repeating the linkage across paraphrases, and displaying calibrated track-records. Because overall liking reduces the gain assigned to new evidence, effective corrections must manage liking and priors as actively as they manage content. For item-level accuracy, deemphasize affect-laden identity markers and, where high liking is expected, plan multi-touch evidence rather than single-shot corrections.

Limitations and future research directions

Our inferences are bounded by design and sample characteristics. The online Australian sample was left-leaning on average and of moderate ideological strength; more moralized or highstakes contexts may produce greater registration of source conflict. The stimulus set (eight leader-attributed statements) limits topic bandwidth and affective variance. Intergroup psychological effects are strongly context-dependent and structured by the current sociopolitical and historical environment. Patterns observed in lowstakes online samples may not generalize to settings where identities are chronically activated and costs are salient (Bilali, 2025).We indexed IED globally and analyzed updating locally, which may understate person-level effects. The mixed-effects model converged with a statement intercept but without a participant intercept. Although we found no evidence of Congruence × IED, our equivalence tests did not establish negligible effects. We expect the observed principle to extend to polarized issues more broadly, but larger and more polarized populations are needed to test this.

A next step is to manipulate and measure RCD directly. For each base claim, researchers can construct graded variants, involving verbatim, paraphrases, referential links, and structured contradictions. Preregister human similarity ratings and embedding-based similarity to anchor these distances. Low RCD should increase coactivation and thus registration of source conflict, making any influence of IED (especially direction-specific IED) observable, particularly on disconfirming trials. High RCD should reduce source-conflict registration.

Conclusion

In this study, fact-checks shifted statement beliefs toward the truth despite a partisan tilt, while conflict in perceived source honesty did not shape that shift. The most coherent account is that conflict was registered at the statement level but remained unregistered at the source level. Advancing corrective interventions will therefore require engineering co activation so that the relevant conflict registers. Because polarization is constructed through cue activation, what is made salient sits upstream of the conflict monitor and influences receptiveness to correction. Identity and source tags can reshape which cues are arbitrated, delaying or weakening local claim–evidence coherence. More broadly, the relationship between conflict, partisanship, and dual-process engagement depends on the intuitions evoked by available cues and prior experience. More research is needed to identify whether conflict results in defensive responses or an openness to adjusting beliefs.

Open Practices Statement

The data, analysis code, materials, and stimuli for this article are available at OSF: https://osf.io/waehy/. This study was not preregistered.

References

Aird, M. J., Ecker, U. K., Swire, B., Berinsky, A. J., & Lewandowsky, S. (2018). Does truth matter to voters? The effects of correcting political misinformation in an Australian sample. *Royal Society open science*, *5*(12), 180593.

Appelman, A., & Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. Journalism & Mass Communication Quarterly, 93(1), 59–79.

Bago, B., & De Neys, W. (2017). Fast logic? Examining the time course assumption of dual process theory. Cognition, 158, 90–109.

Bago, B., & De Neys, W. (2020). Advancing the specification of dual process models of higher cognition: A critical test of the hybrid model view. Thinking & Reasoning, 26(1), 1–30.

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

Bar-Anan, Y. (2020, March 3). Running Project Implicit's IAT from Qualtrics [Blog post]. MinnoJS Blog. https://minnojs.github.io/minnojs-blog/qualtrics-iat/

Baum, J., & Abdel Rahman, R. (2021). Emotional news affects social judgments independent of perceived media credibility. Social Cognitive and Affective Neuroscience, 16(3), 280–291.

Beauvais, N., Voudouri, A., Boissin, E., & De Neys, W. (2025). System 2 and cognitive transparency: deliberation helps to justify sound intuitions during reasoning. Thinking & Reasoning, 1-26.

Bottoms, H. C., Eslick, A. N., & Marsh, E. J. (2010). Memory and the Moses illusion: Failures to detect contradictions with stored knowledge yield negative memorial consequences. Memory, 18(6), 670-678.

Botvinick et al., 2001. Conflict monitoring and cognitive control. Psychological Review, 108(3), 624–652.

Brashier, N. M., & Marsh, E. J. (2020). Judging truth. Annual Review of Psychology, 71, 499–515.

Briñol, P., Petty, R. E., & Wheeler, S. C. (2006). Discrepancies between explicit and implicit self-concepts: Consequences for information processing. Journal of Personality and Social Psychology, 91(1), 154–170.

Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. Journal of Applied Research in Memory and Cognition, 8(1), 108–117.

Bronstein, M. V., Pennycook, G., Joormann, J., Corlett, P. R., & Cannon, T. D. (2019). Dual-process theory, conflict processing, and delusional belief. Clinical psychology review, 72, 101748.

Bullock, J. G. (2009). Partisan bias and the Bayesian ideal in the study of public opinion. Journal of Politics, 71(3), 1109–1124.

Chaiken, S., & Trope, Y. (Eds.). (1999). Dual-process theories in social psychology. Guilford Press.

Chen, E., Chang, H., Rao, A., Lerman, K., Cowan, G., & Ferrara, E. (2021). COVID-19 misinformation and the 2020 US presidential election. Harvard kennedy school misinformation review.

De Neys, W. (2006). Dual processing in reasoning: Two systems but one reasoner. Psychological Science, 17(5), 428–433.

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. Perspectives on Psychological Science, 7(1), 28–38.

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. Thinking & Reasoning, 20(2), 169–187; and De Neys, W. (2023). Advancing theorizing about fast-and-slow thinking. Behavioral and Brain Sciences, 46, e120.

De Neys, W. (2025). Defining deliberation for dual-process models of reasoning. Nature Reviews Psychology, 1-9.

De Neys, W., & Glumicić, T. (2008). Conflict monitoring in dual process theories of thinking. Cognition, 106(3), 1248–1299.

Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. Personality and Social Psychology Review, 14(2), 238–257.

Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. Personality and Social Psychology Review, 14(2), 238-257.

Denes-Raj, V., & Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. Journal of Personality and Social Psychology, 66(5), 819–829.

Dignath, D., Eder, A. B., Steinhauser, M., & Kiesel, A. (2020). Conflict monitoring and the affective-signaling hypothesis—An integrative review. Psychonomic Bulletin & Review, 27(2), 193–216.

Druckman, J. N., & McGrath, M. C. (2019). The evidence for motivated reasoning in climate change preference formation. Nature Climate Change, 9(2), 111–119.

Drummond, C., & Fischhoff, B. (2017). Individuals with greater science literacy and education have more polarized beliefs on controversial science topics. Proceedings of the National Academy of Sciences, 114(36), 9587-9592.

Dujmović, M., & Valerjev, P. (2018). The influence of conflict monitoring on meta-reasoning and response times in a base rate task. Quarterly Journal of Experimental Psychology, 71(12), 2548-2561.

Ecker, U. K., & Ang, L. C. (2019). Political attitudes and the processing of misinformation corrections. Political Psychology, 40(2), 241-260.

Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., ... & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. Nature Reviews Psychology, 1(1), 13-29.

Egner, 2007. Congruency sequence effects and cognitive control. Cognitive, Affective, & Behavioral Neuroscience, 7(4), 380–390.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. Perspectives on psychological science, 8(3), 223-241.

Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. Journal of experimental psychology: general, 144(5), 993.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. P. Zanna (Ed.), Advances in Experimental Social Psychology (Vol. 23, pp. 75–109). Academic Press.

Frey, D., Johnson, E. D., & De Neys, W. (2018). Individual differences in conflict detection during reasoning. Quarterly Journal of Experimental Psychology, 71(5), 1188–1208.

Garcia-Marques, T., Silva, R. R., Reber, R., & Unkelbach, C. (2015). Hearing a statement now and believing the opposite later. Journal of Experimental Social Psychology, 56, 126–129.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. Psychological Bulletin, 132(5), 692–731.

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

Gratton, Coles, & Donchin, 1992. Optimizing the use of information: Strategic control of activation of responses. Journal of Experimental Psychology: General, 121(4), 480–506.

Greenwald, A. G., Brendl, M., Cai, H., Cvencek, D., Dovidio, J. F., Friese, M., ... & Wiers, R. W. (2022). Best research practices for using the Implicit Association Test. Behavior research methods, 54(3), 1161-1180.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. Psychological Science, 17(9), 767–773.

Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. Nature human behaviour, 4(5), 472-480.

Hill, S. J. (2017). Learning together slowly: Bayesian learning about political facts. Journal of Politics, 79(4), 1403–1418.

Jern, A., Chang, K. M., & Kemp, C. (2009). Bayesian belief polarization. Advances in neural information processing systems, 22.

Jern, A., Chang, K.-K., & Kemp, C. (2014). Belief polarization is not always irrational. Psychological Review, 121(2), 206–224.

Johnson, E. D., Tubau, E., & De Neys, W. (2016). The doubting System 1: Evidence for automatic substitution sensitivity. Acta Psychologica, 164, 56–64.

Johnson, I. R., Petty, R. E., Briñol, P., & See, Y. H. M. (2017). Persuasive message scrutiny as a function of implicit–explicit discrepancies in racial attitudes. Journal of Experimental Social Psychology, 70, 222–234.

Kahan, D. M. (2017). Misconceptions, misinformation, and the logic of identity-protective cognition.

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2017). Motivated numeracy and enlightened self-government. Behavioural public policy, 1(1), 54-86.

Karpen, S. C., Jia, L., & Rydell, R. J. (2012). Discrepancies between implicit and explicit attitude measures as an indicator of attitude strength. European Journal of Social Psychology, 42(1), 24–29.

Katabi, N., Simon, H., Yakim, S., Ravreby, I., Ohad, T., & Yeshurun, Y. (2023). Deeper than you think: Partisanship-dependent brain responses in early sensory and motor brain regions. Journal of Neuroscience, 43(6), 1027–1037.

Kemp, P. L., Sinclair, A. H., Adcock, R. A., & Wahlheim, C. N. (2024). Memory and belief updating following complete and partial reminders of fake news. Cognitive Research: Principles and Implications, 9(1), 28.

Kemp, P. L., Wahlheim, C. N., et al. (2022). Recalling fake news during real news corrections can impair or enhance memory updating: The role of recollection-based retrieval. Cognitive Research: Principles and Implications, 7, 61.

Kerns et al., 2004. Anterior cingulate conflict monitoring and adjustments in control. Science, 303(5660), 1023–1026.

Kim, J. W. (2025). Evidence can change partisan minds but less so in hostile contexts. British Journal of Political Science, 55, e62, 1–21.

Leong, Y. C., Chen, J., Willer, R., & Zaki, J. (2020). Conservative and liberal attitudes drive polarized neural responses to political content. Proceedings of the National Academy of Sciences, 117(44), 27731–27739.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. Psychological science in the public interest, 13(3), 106-131.

Lewandowsky, S., Stritzke, W. G., Oberauer, K., & Morales, M. (2005). Memory for fact, fiction, and misinformation: The Iraq War 2003. Psychological Science, 16(3), 190-195.

Luo, Y., & Zhao, J. (2019). Motivated attention in climate change perception and action. Frontiers in Psychology, 10, 1541.

Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. Cognitive Research: Principles and Implications, 5, 47.

McIlhiney, P., Gignac, G. E., Ecker, U. K., Kennedy, B. L., & Weinborn, M. (2023). Executive function and the continued influence of misinformation: A latent-variable analysis. Plos one, 18(4), e0283951.

McLaughlin, B., Holland, D., Thompson, B. A., & Koenig, A. (2020). Emotions and affective polarization: How enthusiasm and anxiety about presidential candidates affect interparty attitudes. American Politics Research, 48(2), 308–316.

Meyer, A., & Frederick, S. (2023). The formation and revision of intuitions. Cognition, 240, 105380.

Nosek, B. A., Bar-Anan, Y., Sriram, N., Axt, J., & Greenwald, A. G. (2014). Understanding and using the Brief Implicit Association Test: Recommended scoring procedures. PLOS ONE, 9(12), e110938.

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(3), 999-1015.

Payne, B. K., Burkley, M., & Stokes, M. B. (2008). Why do implicit and explicit attitude tests diverge? The role of structural fit. Journal of Personality and Social Psychology, 94(1), 16–31.

Pehlivanoglu, D., Lin, T., Deceus, F., Heemskerk, A., Ebner, N. C., & Cahill, B. S. (2021). The role of analytical reasoning and source credibility on the evaluation of real and fake full-length news articles. Cognitive Research: Principles and Implications, 6, 24.

Pennycook, G. (2017). A perspective on the theoretical foundation of dual-process models. In W. De Neys (Ed.), Dual Process Theory 2.0. Routledge.

Pennycook, G. (2023). A framework for understanding reasoning errors: From fake news to climate change and beyond. In Advances in experimental social psychology (Vol. 67, pp. 131-208). Academic Press.

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.

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. Journal of Personality, 88(2), 185–200.

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. Journal of Experimental Psychology: General, 147(12), 1865–1880.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. Cognitive psychology, 80, 34-72.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. Psychological science, 31(7), 770-780.

Pennycook, G., Trippas, D., Handley, S. J., & Thompson, V. A. (2014). Base rates: Both neglected and intuitive. Journal of Experimental Psychology: Learning, Memory, and Cognition, 40, 544–554.

Pennycook, G., Trippas, D., Handley, S. J., & Thompson, V. A. (2014). Base rates: Both neglected and intuitive. Journal of Experimental Psychology: Learning, Memory, and Cognition, 40(2), 544–554.

Petersen, M. B., Giessing, A., & Nielsen, J. (2015). Physiological responses and partisan bias: Beyond self-reported measures of party identification. PLOS ONE, 10(5), e0126922.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), Advances in Experimental Social Psychology (Vol. 19, pp. 123–205). Academic Press.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), Advances in Experimental Social Psychology (Vol. 19, pp. 123–205). Academic Press.

Petty, R. E., Tormala, Z. L., Briñol, P., & Jarvis, W. B. G. (2006). Implicit ambivalence from attitude change. Journal of Personality and Social Psychology, 90(1), 21–41.

Pretus, C., Servin-Barthet, C., Harris, E. A., Brady, W. J., Vilarroya, O., & Van Bavel, J. J. (2023). The role of political devotion in sharing partisan misinformation and resistance to fact-checking. Journal of Experimental Psychology: General, 152(11), 3116.

Puryear, C., Kubin, E., Schein, C., Bigman, Y. E., Ekstrom, P., & Gray, K. (2024). People believe political opponents accept blatant moral wrongs, fueling partisan divides. *PNAS nexus*, *3*(7), page 244.

Rathje, S., Van Bavel, J. J., & Van Der Linden, S. (2021). Out-group animosity drives engagement on social media. Proceedings of the national academy of sciences, 118(26), e2024292118.

Silva, R. R., Garcia-Marques, T., & Reber, R. (2017). The informative value of type of repetition: Perceptual and conceptual fluency influences on judgments of truth. Consciousness and Cognition, 51, 53–67.

Sriram, N., & Greenwald, A. G. (2009). The brief implicit association test. Experimental psychology, 56 (4), 283–294.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. Personality and Social Psychology Review, 8(3), 220–247.

Stump, A., Rummel, J., & Voss, A. (2022). Is it all about the feeling? Affective and (meta-) cognitive mechanisms underlying the truth effect. Psychological Research, 86(1), 12-36.

Su, S. (2022). Updating politicized beliefs: How motivated reasoning contributes to polarization. Journal of Behavioral and Experimental Economics, 96, 101799.

Sultan, M., Tump, A. N., Geers, M., Lorenz-Spreen, P., Herzog, S. M., & Kurvers, R. H. J. M. (2022). Time pressure reduces misinformation discrimination ability but does not alter response bias. Scientific Reports, 12, 22416.

Swire, B., Berinsky, A. J., Lewandowsky, S., & Ecker, U. K. (2017). Processing political misinformation: Comprehending the Trump phenomenon. Royal Society open science, 4(3), 160802.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. American Journal of Political Science, 50(3), 755–769.

Tappin, B. M., & McKay, R. T. (2019). Moral polarization and out-party hostility in the US political context. Journal of Social and Political Psychology, 7(1), 213-245.

Thorson, Emily. "Belief echoes: The persistent effects of corrected misinformation." Political Communication 33.3 (2016): 460-480.

Unkelbach, C., & Greifeneder, R. (2013). A general model of fluency effects in judgment and decision making. In C. Unkelbach & R. Greifeneder (Eds.), The Experience of Thinking (pp. 21–34). Psychology Press.

Unkelbach, C., & Greifeneder, R. (2013). A general model of fluency effects in judgment and decision making. In The experience of thinking (pp. 11-32). Psychology Press.

Unkelbach, C., & Rom, S. C. (2017). A referential theory of the repetition-induced truth effect. Cognition, 160, 110–126.

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

Van Boven, L., Ramos, J., Montal-Rosenberg, R., Kogut, T., Sherman, D. K., & Slovic, P. (2019). It depends: Partisan evaluation of conditional probability importance. Cognition, 188, 51–63.

van Harreveld, F., Nohlen, H. U., & Schneider, I. K. (2015). The ABC of ambivalence: Affective, behavioral, and cognitive consequences of attitudinal conflict. Advances in Experimental Social Psychology, 52, 285–324.

Walter, N., & Tukachinsky, R. (2020). A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it?. Communication research, 47(2), 155-177.

Walter, N., & Tukachinsky, R. (2020). A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it. Communication Research, 47(2), 155–177.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. Psychological Review, 107(1), 101–126.

Wischnewski, M., & Krämer, N. (2021). The role of emotions and identityprotection cognition when processing (mis)information. Technology, Mind, and Behavior, 2(1).

Yu, S., & Opfer, J. E. (2024). Cognitive support for political partisans’ understanding of policy data. PLOS ONE, 19(10), e0312088.

Zhang, Y., & Rand, D. G. (2023). Sincere or motivated? Partisan bias in advice-taking. Judgment and Decision making, 18, e29.

Zlotnick, E., Dzikiewicz, A. J., & Bar-Anan, Y. (2015). Minno.js (Version 0.3) [Computer software]