Simple Logic Warnings Improve Climate Argumentation Across Prior Beliefs and Political Identities

Bence Bago[1] & Zoe A. Purcell[2]

1. Department of Social Psychology, Tilburg University, Tilburg, The Netherlands
2. LaPsyDÉ, Université Paris Cité, CNRS, F-75005 Paris, France

Corresponding Author: Bence Bago, Department of Social Psychology, Tilburg University, e.mail:b.bago@tilburguniversity.edu

Abstract

Collective climate action requires shared climate beliefs in line with the scientific consensus. However, climate discourse is flooded with fallacious information and people often fail to evaluate it correctly, accepting weak arguments and rejecting strong ones. Understanding these failures is critical for developing evidence-based interventions, yet it remains unclear whether they are driven by partisan identity protective tendencies, entrenched prior beliefs, or failures in logical reasoning. This opacity stems from a persistent empirical challenge: realistic climate argumentation inherently mixes these factors together, making it difficult to isolate their independent effects without sacrificing ecological validity. Across three preregistered studies (N = 3,000), we introduce a protocol that resolves this trade-off by orthogonally manipulating argument logicality and direction across realistic contexts. In neutral contexts, prior beliefs dominated judgments, with smaller effects of partisanship and logicality. Contrary to identity-protective accounts, embedding arguments in hostile partisan contexts did not increase partisan bias or impair logical reasoning. Instead, partisan attacks reduced reliance on prior beliefs, prompting participants to downgrade belief-consistent arguments. In a third study, minimal logic warnings—brief alerts indicating possible fallacies—enhanced discrimination between weak and strong arguments. These findings challenge single-factor accounts of climate argumentation, showing that argumentation is context sensitive and that this sensitivity can be exploited to improve climate argumentation by adding simple logic warnings that are effective, scalable, and politically neutral.

Introduction

Addressing the climate crisis requires a united front. However, despite the overwhelming scientific consensus on climate change, public opinion remains divided. The digital information ecosystem is saturated with logically flawed anti-climate arguments that continue to gain traction despite their evident weaknesses [1–4]. Why do people fail to discard these arguments? And can we improve it?

Theories of *partisan identity protection* suggest that people prefer their beliefs to be aligned to that of their partisan group, and evaluate them based on this alignment, regardless of their logical flaws[5–16]. Theories on *coherence-based* reasoning suggest that people prefer not to change their mind, and evaluate arguments based on whether they support their prior beliefs, again regardless of their logical flaws[17–25]—arguments that align with existing views are more likely to be accepted, while conflicting ones are often rejected. These theories imply that logical reasoning abilities matter little when people evaluate arguments on politically controversial or topics they have a strong opinion on.

This is at odds with results showing that people not only have the capacity to reason well, but also reason better when encouraged. Indeed, studies on *logical reasoning* show that people are sensitive to the logical quality of arguments—their internal consistency, valid inference, and evidential support [26–29]. People are not only capable of logical reasoning when thinking effortfully but also under strict time-limits and in low-effort contexts such as social media where much of the climate discourse takes place [29–32]. Unlike identity protection and coherence-based reasoning, logical reasoning has been shown to improve (increasing responses in line with normative, logical, and probabilistic rules) with very simple training interventions in non-politicized contexts [33–35].

Yet, there are practical reasons why research on climate communication has not fully tackled the challenge of untangling the respective direct and indirect effects of partisan identity (as per identity protection accounts), prior beliefs (as per coherence-based accounts) and logical quality (as per logical reasoning accounts).

First, in the case of climate arguments—these factors naturally covary. In the United States, partisanship correlates with climate beliefs (i.e., Republicans are less likely than Democrats to accept climate science), confounding identity with priors. Furthermore, anti-climate arguments are more frequently logically flawed than pro-climate arguments, creating a confound between argument direction and logical quality. Pro-climate arguments tend to maintain logical coherence, whereas anti-climate arguments frequently rely on reasoning fallacies such as false dichotomies, circular reasoning, or emotionally charged appeals [36–38]. Consequently, existing studies—predominantly correlational and without rigorous control—cannot isolate the independent contributions of these factors nor establish causal relationships.

Second, experimental manipulations designed to examine causality often do so in isolation (considering only one factor) and using unrealistic manipulations. Studies designed to activate identity-protective reasoning—such as instructing participants to "view this problem through a political lens"—are highly contrived and may not generalize to naturalistic contexts [39,40]. Moreover, such manipulations may not engage the intended psychological processes: viewing issues politically does not necessarily threaten one's identity, leaving open whether observed effects reflect identity protection or other mechanisms. Interventions designed to improve logical reasoning are similarly limited to contrived, lab-based settings that, while theoretically informative, may not extend to relevant types of arguments and fallacies that plague climate argumentation in the wild [33–35]. Thus, while evidence suggests that partisanship, priors, and logic may each influence climate argumentation, their relative contributions, causal roles, and interplay in realistic, variable contexts remain unknown.

To overcome these challenges, we developed a protocol in which we tightly control each factor within realistic experimental contexts. Participants first report their partisan identity and their specific climate beliefs regarding its existence, anthropogenic cause, and potential threat. They then evaluate the strength of 40+ arguments that are fully crossed in direction (pro- or anti-climate arguments), topic (existence, anthropogenic cause, or potential threat), and logical quality (with or without a logical fallacy). Using this protocol, Study 1 established the baseline contributions of identity, priors, and logic in a 'neutral context'—without additional partisan or logical cues. Study 2 tested whether embedding these arguments in an ‘identity-threatening context’—emulating politicised social media discourse by adding partisan ad hominem attacks in the form of a 'retweet'—amplified reliance on partisanship. Finally, Study 3 examined whether arguments placed in a ‘logic-salient context’ improved logical reasoning. Reflecting realistic social media discourse once again, arguments containing fallacies were accompanied by ‘warnings’ that a fallacy may be present.

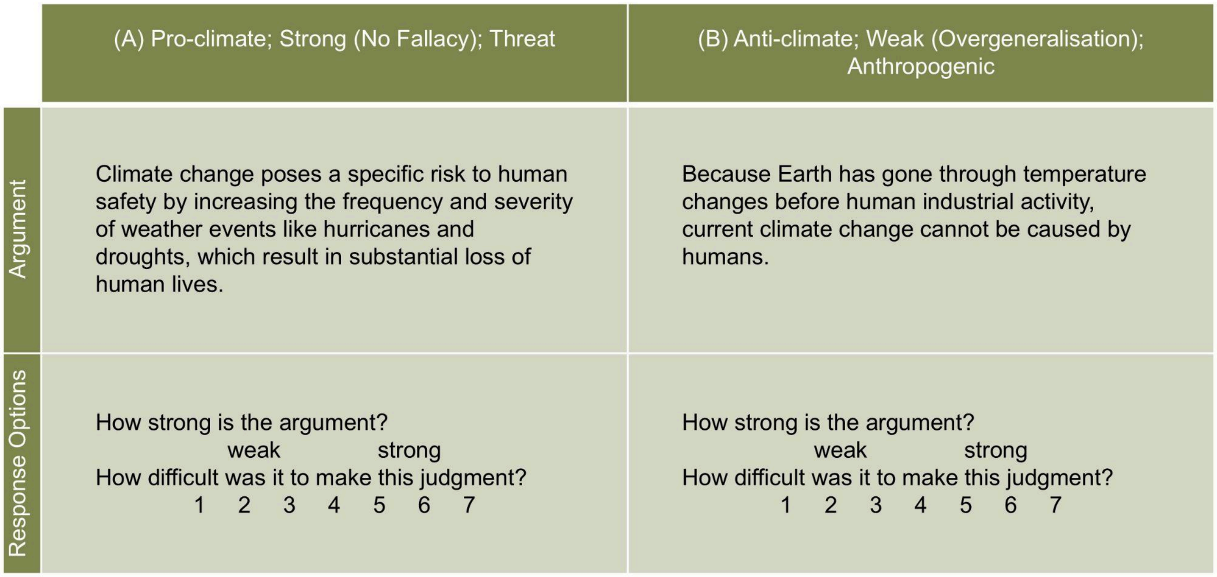
These three cumulative pre-registered studies overcome issues of natural covariation and unrealistic manipulations, demonstrating that (1) prior beliefs dominate climate argumentation across contexts, (2) partisanship plays a limited role and one that runs counter to prominent identity-protective accounts, and (3) in contrast to both identity-protective and coherence-based accounts—people's ability to discern weak from strong arguments is *not* jeopardised by either prior beliefs or partisan motivations. In this picture, logic is painted the dark horse—its consistent impact across contexts, capacity for improvement via simple prompts, and natural covariation with arguments in line with the scientific consensus points to an overlooked, actionable opportunity for improving climate argumentation.

Paradigm

We examined how people evaluate climate change arguments across three experiments (sample quota-matched for age, gender and political affiliation, collected through Lucid; total N = 3000) using a novel paradigm that independently manipulated argument logic, stance (pro- or contra-climate change), and contextual framing. Participants evaluated arguments containing either logical fallacies (post hoc reasoning, denying the antecedent, affirming the consequent, or overgeneralization) or sound logical structures. For each argument (48 in Study 1, 32 in Study 2 and 3 - to fit in 10 minutes survey), participants had to answer 'How strong is the argument?' by clicking on 'Strong' or 'Weak'. To most rigorously assess the effect of priors, we matched specific prior belief measures and argument content across three dimensions: climate change existence, anthropogenic causation, and risk severity. For example, arguments about the risks of climate change were matched with each participants' prior on risk. As a control, and to provide variety, maintain the attention of our participants, and limit experimenter bias, we also measured how people evaluate arguments on less politicized topics including Genetically Modified Organisms (GMO), nanotechnology, and Artificial Intelligence (AI).

Results

Study 1: Prior beliefs dominate climate argument evaluation in a neutral context



	(A) Pro-climate; Strong (No Fallacy); Threat	(B) Anti-climate; Weak (Overgeneralisation); Anthropogenic
Argument	Climate change poses a specific risk to human safety by increasing the frequency and severity of weather events like hurricanes and droughts, which result in substantial loss of human lives.	Because Earth has gone through temperature changes before human industrial activity, current climate change cannot be caused by humans.
Response Options	How strong is the argument? weak strong How difficult was it to make this judgment? 1 2 3 4 5 6 7	How strong is the argument? weak strong How difficult was it to make this judgment? 1 2 3 4 5 6 7

Figure 1. *Study 1 presented arguments in a neutral context, without additional political or logical cues. On the left is an example of an argument that is pro-climate, strong, and about the threat posed by climate change. On the right is an example of an argument that is anti-climate, weak, and about the anthropogenic nature of climate change.*

In Study 1, prior beliefs emerged as the dominant factor impacting the judgment of arguments as weak or strong, consistently with coherence-based reasoning accounts. We first assessed the direct effects of priors, logic, and partisanship. Priors had a significant effect: participants rated belief-consistent arguments substantially stronger than belief-inconsistent ones, *b* = 2.78, *p* < .001. We also found smaller but significant effects of logic, *b* = 0.40, *p* = .008, and partisan identity, *b* = 0.81, *p* < .001: logical arguments and those aligned with one's politics tended to be rated as strong more often than arguments with fallacies or that did not align with one's politics.

We then explored how these factors interacted to affect argumentation. In particular, we analysed heterogeneity of the effect of logic on argument evaluation. That is, whether the effect of logic varied due to interference created by priors or politics. We found that logic, priors, and argument direction interacted to significantly impact judgments, *b* = 0.77, *p* < .001. Post hoc analyses suggested that priors interfered: people were better at discerning weak from strong arguments when the argument direction was consistent with people's priors, *b* = 0.65, *p* < .001, versus when it was not, *b* = 0.36, *p* = .014. However, we found no evidence that politics interfered with logic: the three-way logic, politics, and argument direction interaction was not significant, *b* = -0.01, *p* = .949. See SI Table S1 for extended results. For non-politicised arguments (e.g., about GMO), a similar pattern emerged with regard to priors and logic, however, as expected, politics had neither a direct nor an interference effect, see SI Tables S3-4 for details.

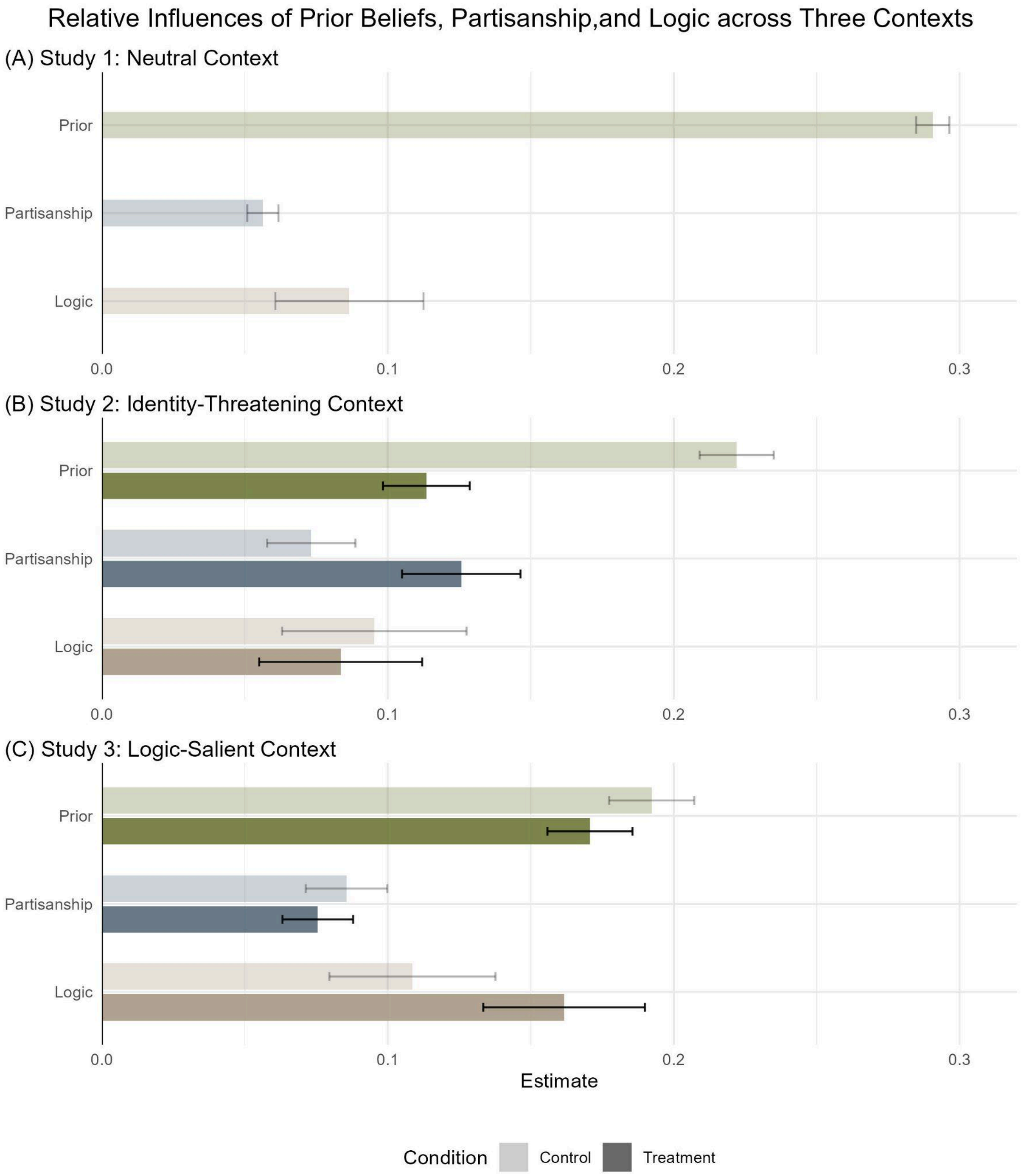
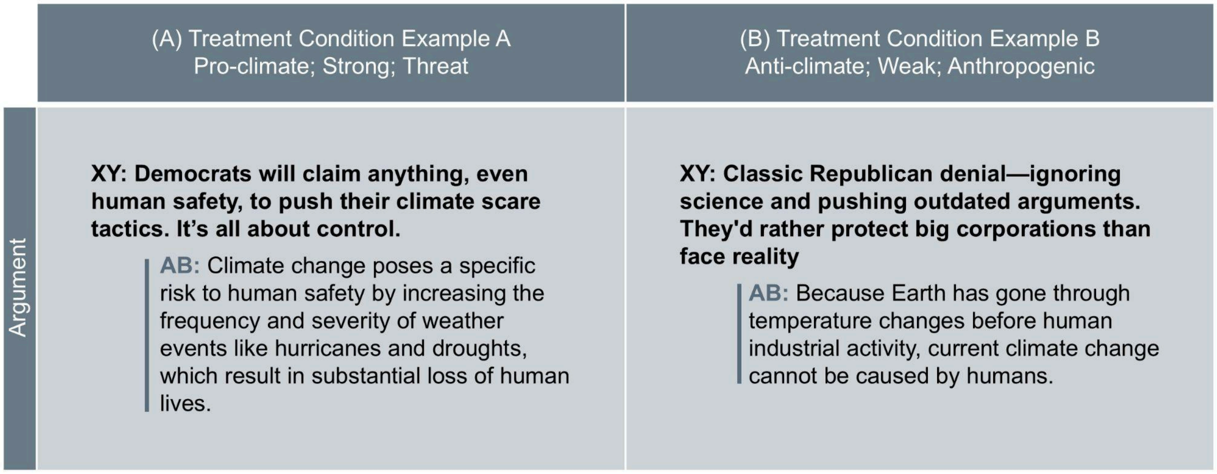


Figure 2. *In a neutral context (A), participants' prior beliefs were the strongest predictor of their evaluation of argument strength. In identity-threatening contexts (B), the influence of prior beliefs decreased and the influence of political ideology increased—but not in the expected direction (see main text). In logic-salient contexts (C), participants were more sensitive to the logical quality of the argument. Note: For comparability, standardized beta coefficients are presented with 95% confidence intervals.*

Study 2: Partisan attacks paradoxically reduce polarization in an identity-threatening context

To examine whether the identity causally impacts argument evaluation, we developed a controlled manipulation designed to approximate key features of social media while maximizing identity threat. Participants were randomly allocated to the political ‘treatment condition’—viewing the climate arguments alongside hostile partisan attacks in a quote share format—or the ‘control condition’ —viewing the arguments in isolation, as in Study 1. The partisan attacks targeted Democrats for anti-climate arguments(see Figure 3A) and Republicans when accompanying pro-climate arguments (see Figure 3B), creating maximum threat to participants' political identity.



	(A) Treatment Condition Example A Pro-climate; Strong; Threat	(B) Treatment Condition Example B Anti-climate; Weak; Anthropogenic
Argument	**XY: Democrats will claim anything, even human safety, to push their climate scare tactics. It’s all about control.** AB: Climate change poses a specific risk to human safety by increasing the frequency and severity of weather events like hurricanes and droughts, which result in substantial loss of human lives.	**XY: Classic Republican denial—ignoring science and pushing outdated arguments. They'd rather protect big corporations than face reality** AB: Because Earth has gone through temperature changes before human industrial activity, current climate change cannot be caused by humans.

Figure 3. *In Study 2, participants were placed in either the treatment or control condition. Those in the treatment condition evaluated arguments both with and without a partisan attack, while those in the control condition were not exposed to partisan attacks. On the left is an example of a pro-climate argument treated with an anti-democrat partisan attack. On the right is an example of an anti-climate argument treated with an anti-republican attack.*

This manipulation provides a realistic, direct, and rigorous test of the role of identity in argument evaluation. First, the quote-share format mirrors real-world social media contexts where climate arguments are frequently embedded in partisan rhetoric [2–4]; precisely the contexts where identity-protective reasoning is theorized to be most influential [2,11,41,42]. Second, by pairing arguments with ad hominem attacks that explicitly target participants' political group, we directly manipulate identity threat independent of argument content. This addresses a fundamental confound in prior research: climate arguments are inherently identity-threatening to some groups (e.g., pro-climate arguments threaten conservative identity), but this identity threat is confounded with the argument's consistency with prior beliefs, making it impossible to isolate the causal role of identity from belief-based reasoning. Our manipulation disentangles these factors—by adding explicit identity attacks on top of the argument, we can observe whether activating political identity leads to defensive reasoning. Third, the attacks are severe and direct. They do not present rational counter-arguments but personally denigrate the participant's political group, creating a context where identity defense mechanisms should be maximally activated if they operate as theorized.

If political identity affects argument evaluation, this could occur through two theoretically distinct pathways. First, identity could operate directly by motivating people to defend their political group regardless of their substantive beliefs about climate [43–45]. Under this account, identity threat should increase the direct effect of partisanship on argument evaluation, and people should more strongly endorse arguments that align with their partisan group and reject those that don't, independent of their actual beliefs about climate change. Second, identity could operate indirectly by strengthening reliance on prior beliefs that are often aligned with one's political group [17,25]. Under this account, identity threat should amplify the effect of priors—people become more certain in beliefs that define their political identity when that identity is challenged.

Despite these mechanistic differences, both pathways converge on a core prediction: if people engage in identity-protective reasoning, they should reinforce their existing beliefs when their identity is threatened. That is, people should evaluate belief-consistent (or identity-consistent) arguments as strong and belief-inconsistent arguments as weak when their identity is threatened compared to neutral contexts [13,40]. Study 2 directly tests this prediction.

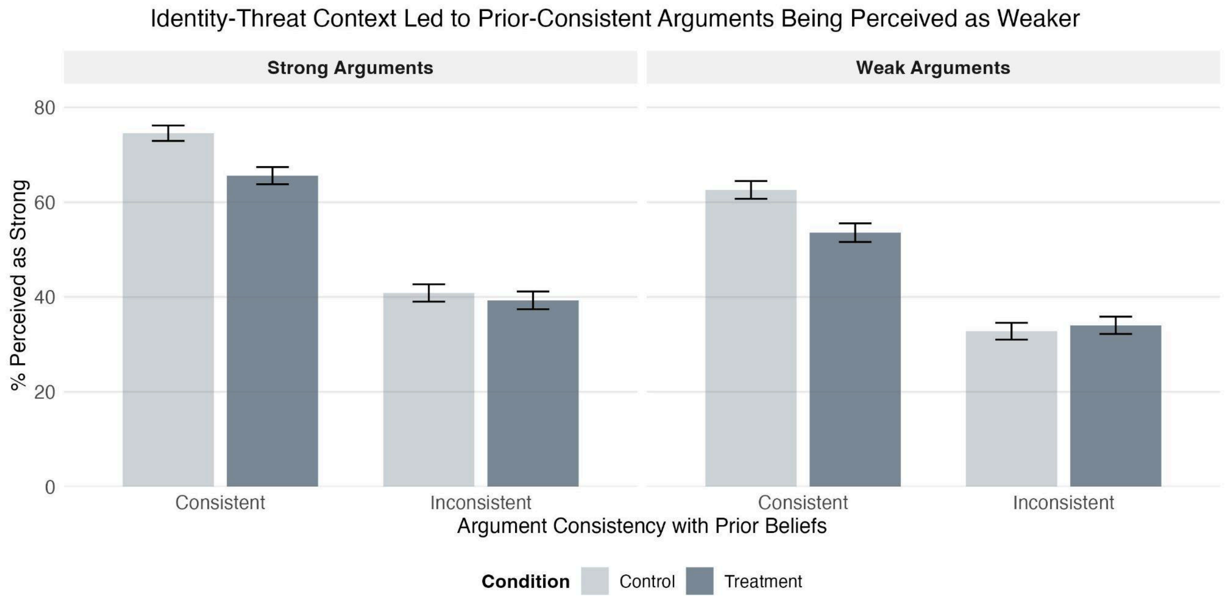


Figure 4. *In Study2, the percent of people perceiving an argument as strong was affected by their prior beliefs and the argument's logicality. Regardless of actual argument strength, consistent arguments presented in politicised contexts (dark bars) were perceived as strong more often than consistent arguments presented in neutral contexts (light bars). Error bars are 95% confidence intervals.*

We first analysed the direct effects of priors, partisanship, logic, and treatment. As in Study 1, prior beliefs had the largest effect on argument evaluation, *b* = 2.31, *p* < .001, followed by partisan identity, *b* = 0.68, *p* < .001. The main effect of logic was similar to that observed in Study 1, under neutral conditions (*b* = 0.30, *p* = .080; standardised direct effects are presented in Figure 2). These results replicate the pattern from Study 1, again suggesting that prior beliefs play a substantially larger role than political identity in shaping argument evaluation—a pattern that poses challenges for strong versions of identity-protective accounts that treat partisan identity as the primary driver of argument evaluation on climate change.

We then examined whether the identity treatment interfered with these effects. Counter to identity-protective predictions, we found no evidence that political identity influence increased under partisan attack. The effect of partisanship did not significantly increase under treatment (treatment × partisanship × argument direction: *b* = 0.13, *p* = .463). The effect of logic (i.e., people's ability to discern fallacious from valid arguments) was not directly impacted by the treatment either (treatment × logic: *b* = 0.02, *p* = .807). This stable effect of logic under partisan attack further challenges theories suggesting that threats to political identity necessarily harm people's ability to think clearly.

More strikingly, the influence of prior beliefs actually decreased under identity attack (treatment × prior belief × argument direction: *b* = -0.97, *p* < .001). This pattern is the opposite of what identity-protective accounts predict—rather than becoming more entrenched in their existing beliefs when faced with identity attacks, participants became less convinced by arguments consistent with their priors. As depicted in Figure 4, post-hoc simple effect analyses revealed that participants actually downgraded belief-consistent arguments when wrapped in a political context, *b* = -0.54, *p* < .001, while they maintained their evaluation of belief-inconsistent arguments, *b* = -0.01, *p* = .912. This asymmetry held across the political spectrum, both for Democrats and Republicans (but effect was smaller for Democrats), climate skeptics and believers (see SI). The fact that treatment impacted belief-consistent but not inconsistent arguments, suggests that the effect is not a general devaluation of all arguments paired with attacks. This pattern of results suggests there is neither a direct nor indirect identity-protective mechanism.

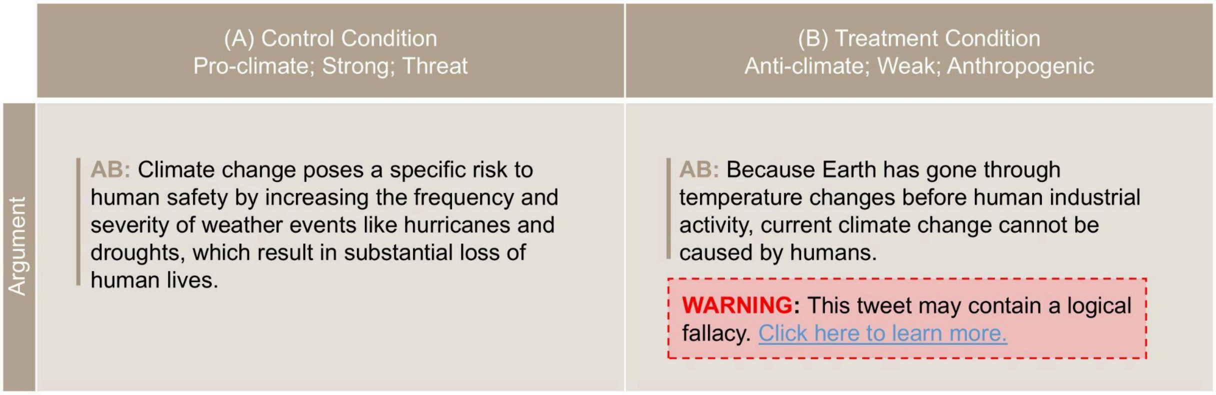
What may cause people to devalue belief-consistent arguments when accompanied with identity-threatening comments? There are two distinct possibilities: people may be responding to its identity-threatening nature or they may be reacting to the disagreeing nature of it (i.e., that some of their peers may disagree with such opinions). To disentangle these options, we look to the non-climate arguments. Recall that participants were also presented with arguments about politically neutral scientific topics (i.e., about GMO, nanotechnology, and AI). For those in the treatment condition, we accompanied these items with attacks that do not mention politics but simply disagree with the content of the target argument. For example, the following target argument: *'Claiming all GMO food is risky for consumers ignores the diverse range of GMO crops and their individual safety assessments.'* was accompanied with the following retweet: '*Ah, so they want us to believe GMOs are harmless because they're 'different'? Sounds like someone's looking for excuses to ignore real concerns about what we're putting on our plates.*'.

We conducted the same analysis on non-climate arguments as we did for climate arguments and observed that these retweets produced the same pattern as partisan attacks: decreasing reliance on prior beliefs for evaluating arguments, treatment x prior beliefs x argument direction, *b* = -0.53, *p* = .020. The fact that mere disagreement—even without identity threat—produces the same effect suggests that what matters is not the identity-attacking nature of the retweet but rather that it signals controversy and contested beliefs. This supports coherence-based rather than identity-protective accounts: people appear to update their evaluations when they encounter signals that others disagree, regardless of whether those disagreements invoke political identity. Extended results for Study 2, including those on non-climate items can be found in the SI Tables S5-8.

Finally, as for Study 1, we also examined the heterogeneity of the effect of logic on argument evaluation and whether people's ability to distinguish strong from weak arguments varied due to interference created by priors or politics. We observed little heterogeneity in the effect of logic due to partisanship or priors. The three-way prior x argument direction x logic effect was not significant, *b* = -0.104, *p* = .663, suggesting the logic effect did not differ between prior-consistent and prior-inconsistent arguments. Moreover, the four-way treatment x prior x argument direction x logic effect was not significant, *b* = 0.31, *p* = .344, suggesting that the logic effect was maintained despite variation in both treatment and the argument's consistency with priors. Similarly, we observed little heterogeneity in the effect of logic across argument consistency with partisanship (partisanship x argument direction x argument strength, *b* = 0.12, *p* = .656) and treatment (treatment x partisanship x argument direction x argument strength, *b* = 0.23, *p* = .507).

Logic-salient context: Minimal warnings increase logical responding

Study 1 and 2 showed a consistent but weak effect of logic. This suggest that people are sensitive to logic, but some of them might lack the mindware to recognize logical fallacies or they might underweight it compared to other cues. To solve this, in Study 3, we introduced a simple intervention to increase logic salience. The design was motivated by fact-checking interventions on social media platforms [46–48], which typically present minimal initial information—such as flags indicating disputed content—and require users to actively click for detailed explanations. However, rather than fact-checking, we implement *logic-checking*—a strategy that is more scalable and places less pressure on platforms to adjudicate politically sensitive claims, as it focuses on the structure of arguments rather than their factual or ideological content. The current two-step design mirrors real-world information ecosystems and allows us to examine both passive exposure to warnings and active engagement with logical explanations.


(A) Control Condition
Pro-climate; Strong; Threat
(B) Treatment Condition
Anti-climate; Weak; Anthropogenic
Argument

AB: Climate change poses a specific risk to human safety by increasing the frequency and severity of weather events like hurricanes and droughts, which result in substantial loss of human lives.

AB: Because Earth has gone through temperature changes before human industrial activity, current climate change cannot be caused by humans.

WARNING: This tweet may contain a logical fallacy. Click here to learn more.

Figure 5. *Study 3 presented arguments either with or without a warning that there may be a logical fallacy in the argument. Those in the treatment condition saw strong arguments without a warning (as in A) and weak arguments with a warning (as in B). Those in the control condition did not see any warnings (as in A). On the left is an example from the strong items which were not presented with warnings. On the right is an example from the weak arguments which, for those in the treatment condition, were presented with warnings. When presented with a warning, participants had the option to follow the link and learn more about the specific fallacy.*

As depicted in Figure 5B, we applied a 'warning treatment'—AI-generated alerts appeared on half of the arguments containing fallacies. This initial warning provided minimal information, signaling potential logical problems without specifying the nature of the flaw. Participants could then click on the warning to receive an additional explanation about the specific fallacy detected in the argument. For example, "*The post hoc ergo propter hoc fallacy was identified. Post hoc ergo propter hoc fallacy means assuming that because one event follows another, the first event caused the second. This reasoning is invalid because volcanic eruptions causing climate changes in the past do not disprove human-induced climate change.*". This format enhances ecological validity: like real social media fact-checks, our intervention requires minimal effort to notice but offers deeper engagement for those seeking details. The minimal intervention was deliberately designed to test whether even brief attention to logical structure—without extensive training or elaborate interventions—could shift argument evaluation.

The design tests a specific mechanism: if logic's modest influence reflects limited salience (or lack of knowledge) rather than motivated resistance, then making logical flaws explicit should enhance their impact on evaluation. The critical question is whether this effect operates uniformly—if warnings enhance logical discernment regardless of belief consistency, it would support the view that logic is under-weighted but intact in politicized reasoning. However, if prior beliefs or partisan identity moderate the effect of warnings, with people selectively ignoring logical flaws in belief-consistent arguments, it would suggest deeper motivated barriers to logical evaluation. This would demonstrate that the resilience of logic observed in Studies 1 and 2 reflects only passive discernment, not active responsiveness to logical cues when they conflict with preferred conclusions.

We first analysed the direct effects of priors, partisanship, and logic, and treatment. As in Studies 1 and 2, prior beliefs remained the strongest predictor of argument evaluation, $b = 2.36, p < .001$, while logic, $b = 0.50, p = .002$, and partisan identity showed smaller but significant effects, $b = 0.88, p < .001$. The warning treatment had no significant main effect on perceived argument strength, $b = -0.16, p = .121$, indicating that warnings did not simply increase overall skepticism toward all arguments. Rather, as we explore below, the treatment selectively enhanced people's ability to discriminate between logically valid from invalid arguments.

We then examined whether the logical warnings interfered with the influence of partisanship, priors, and logic on argument evaluation (see Figure 2). The treatment did not significantly decrease the effect of politics (treatment × partisanship × argument direction: $b = -0.10, p = .494$). Partisan identity maintained its influence across both control and treatment conditions, suggesting that while logical warnings enhanced discernment, they did not override or substantially reduce identity-based reasoning. This pattern indicates that logical cues and partisan considerations operate as parallel—rather than competing—processes in argument evaluation.

The intervention significantly decreased people's reliance on prior beliefs when evaluating arguments (treatment × prior beliefs × argument direction: $b = -0.43, p = .002$). This effect mirrors the pattern observed in Study 2, where encountering signals of disagreement (in that case, hostile partisan attacks) led people to question their conviction in belief-consistent arguments. Here, explicit warnings about logical flaws appear to serve a similar function: they prompt reconsideration of whether arguments aligning with one's priors are genuinely strong. This reduction in prior-based reasoning suggests that logical warnings not only enhance attention to argument structure but also moderate the dominant influence of prior beliefs—a finding with important implications for interventions aimed at promoting more balanced argument evaluation. That is, logic interventions can serve as a precursor for belief change: relying less on one's original priors when evaluating arguments could help people to build new, more accurate beliefs.

Critically, the intervention successfully enhanced discrimination between logically valid and invalid arguments. The treatment increased logical discernment (treatment × logic: $b = 0.28, p = .003$), driven by a 4 percentage point increase in judging fallacious arguments as weak, and a 1 percentage point increase in judging strong arguments as strong. At scale, this effect represents a substantial improvement in people's ability to distinguish strong from weak arguments based on their logical structure, demonstrating that even minimal interventions can meaningfully shift how people evaluate arguments about climate change (Figure 6).

Crucially, these logical improvements occurred uniformly across prior beliefs and partisan identities. The intervention's effectiveness did not vary with belief consistency (treatment × prior beliefs × argument direction × logic: $b = 0.05$, $p = .840$). Post-hoc analyses confirmed that the intervention improved argument discernment both when arguments were consistent with one's priors, $b = 0.33$, $p < .001$, and when they were inconsistent, $b = 0.24$, $p < .001$. However, the effect on inconsistent arguments was driven by higher acceptance of strong arguments, while the effect on consistent arguments was driven by a decrease in the acceptance of weak arguments (see SI). Similarly, the intervention worked uniformly across partisan alignment (treatment × partisanship × argument direction × logic: $b = 0.02$, $p = .940$).

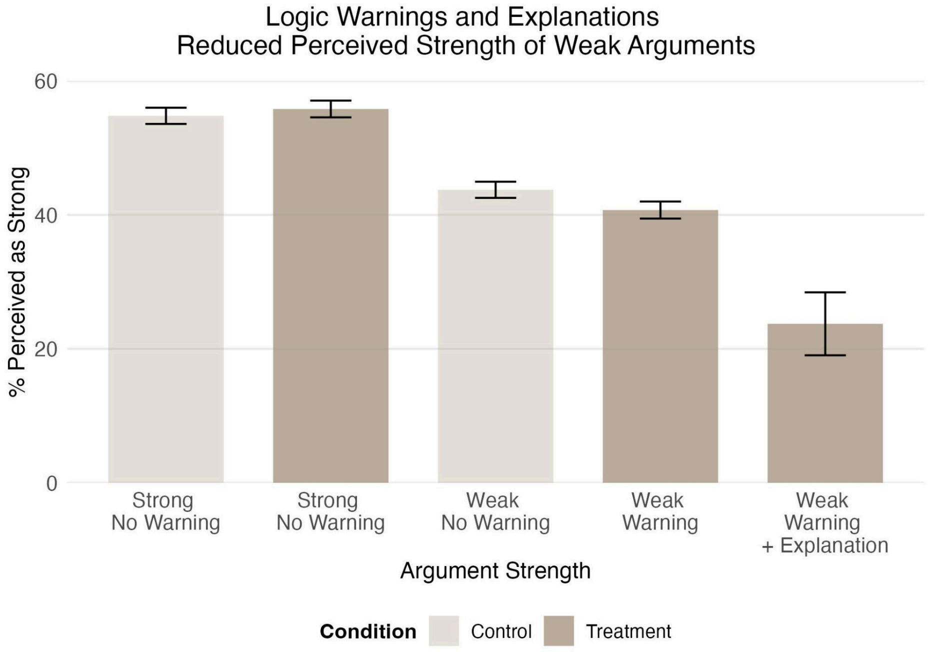


Figure 6. *Logical warnings increase the ability to discern weak from strong arguments. Figure shows perceived strength (percent of people perceiving the argument as strong) per argument's consistency with prior beliefs and their strength (Strong = argument has no fallacy, weak = argumentative fallacy). Orange bars are averages for logical context, in which target weak arguments were accompanied with a warning that the argument may contain a logical mistake, and blue bars are averages for the control condition in which arguments were presented on their own. Error bars are 95% confidence intervals.*

To test the generalizability of these effects beyond climate change, we replicated the analysis on politically neutral items (arguments about GMO, nanotechnology, and AI). The treatment was equally effective on neutral items, producing a 5% increase in discernment (treatment × logic: $b = 0.29$, $p = .009$). This replication demonstrates that the effectiveness of logical warnings is not specific to climate arguments but reflects a general mechanism applicable across different domains of argumentation.

On climate arguments with warnings, participants clicked on the link to learn more about the specific logical fallacy only 5% of the time, suggesting that most people do not actively seek out additional logical information. One might assume, however, that the effect may be even stronger for those who engage with the detailed explanations. To explore this, we ran an exploratory analysis excluding weak arguments from the treatment condition when the button was not clicked; this led to an increase in discernment among weak arguments to 21%. While we cannot draw causal conclusions regarding the effect of the information in the popup window due to self-selection, this result is consistent with previous studies on fact-checking suggesting that warnings are most effective when accompanied by specific, content-relevant explanations [49]. This pattern suggests that while passive exposure to warnings produces meaningful effects, interventions may be further optimized by encouraging or automatically providing detailed logical explanations alongside initial warnings. Extended results for Study 3 climate and neutral items can be found in the SI.

Discussion

A major impediment to collective climate action is divided climate beliefs. In an increasingly complex information environment, people's ability to discern strong from weak arguments is a critical component of forming accurate climate beliefs. In this article, we explored what causes poor argument evaluation. In three studies, our findings consistently challenge the prominent identity-protective accounts, illustrate the importance of prior beliefs, and demonstrate that even when people have different political affiliations and prior beliefs—they remain sensitive to the logical quality of the argument. These theoretical observations directly informed the development and testing of a successful and scalable logic-based intervention that improves climate argumentation.

A central question motivating this research was whether identity-protective cognition causally drives climate argument evaluation. Identity-protective theory predicts that threats to one's political identity should trigger defensive reasoning—strengthening endorsement of identity-consistent arguments and increasing rejection of opposing ones. Study 2 explicitly maximized this threat: hostile partisan attacks directly denigrated participants' political groups in a format mirroring real social media contexts where identity protection should be most active.

Despite this maximal design, we found no causal evidence supporting the identity-protective account. Partisan attacks made people *less* confident in belief-consistent arguments while leaving belief-inconsistent arguments unchanged—the opposite of identity-protective predictions. This pattern held uniformly across Democrats and Republicans and generalized to non-political topics (GMOs, nanotechnology, AI) when paired with disagreeing comments containing no identity threat. If political identity were the mechanism, effects should have emerged only for climate arguments where partisan stakes are high. However, mere disagreement—without identity threat—produced the same belief moderation demonstrating that the mechanism is not identity-based.

In line with previous studies, political identity exerted a direct effect on argument evaluation—people tended to rate arguments aligned with their group as stronger. This pattern has often been taken as evidence that identity, particularly political identity, impairs sound argumentation. Yet, our findings suggest this conclusion may have been premature. Inducing identity threat through hostile attacks neither amplified partisan responding nor disrupted logical discernment. Instead, the results indicate that partisanship shapes baseline argumentation but does not operate through identity-defensive mechanisms that undermine reasoning.

What, then, drives (poor) climate argument evaluation? The observed inverse effect is more consistent with coherence-based accounts. Hostile disagreement appears to signal that an issue is contested, prompting individuals to reconsider the certainty of their prior beliefs–even when the disagreement targets their identity. Rather than reflexively defending their position, participants seem to engage in belief revision, reconsidering whether belief-consistent arguments are as strong as they may have initially assumed.

This pattern also aligns with contemporary dual-process models of reasoning, which distinguish between automatic and effortful cognitive processes and emphasise the role of metacognitive signals—such as uncertainty—in regulating transitions between them [26,29,50–52]. Judgments anchored in prior beliefs may arise from relatively automatic, intuitive processing, but encountering disagreement can function as a metacognitive cue that initiates more effortful, reflective evaluation. When contextual cues highlight controversy or disagreement, individuals may temper their confidence in prior-consistent arguments without correspondingly increasing their endorsement of prior-inconsistent ones, toward which they were already skeptical. This asymmetric updating—downgrading accepted claims when contested but not upgrading rejected claims—suggests rational belief adjustment rather than defensive reinforcement.

Prior beliefs thus appear to operate as a context-sensitive baseline: dominant in neutral settings but moderated when environmental signals indicate controversy (partisan attacks) or highlight structural quality (logic warnings). Political identity operates as a stable preference that does not escalate defensively under threat and logic operates in a preserved but under-weighted capacity. That is, people distinguish valid from fallacious arguments across all contexts—neutral, identity-threatening, belief-consistent, and belief-inconsistent—but this capacity exerts weak influence unless explicitly highlighted. When logic is explicitly highlighted—as when logic warnings were applied in Study 3—its impact is uniformly increased, demonstrating that the initially modest effect of logic influence may reflect limited salience or recognition rather than motivated resistance.

These findings reject single-factor explanations and challenge the prevailing emphasis on identity-protective cognition. Climate argument evaluation reflects context-sensitive reasoning where multiple factors operate in parallel, shifting in relative influence as contextual cues change [26,51–53]. The critical question that emerges is not which factor dominates, but under what conditions each gains or loses influence.

Practical Implications

While our primary contribution is theoretical, the findings also inform intervention development. Based on theoretical findings, we developed and tested a logic-based treatment—observing improved argumentation across priors and partisanship. This approach offers several advantages. First, logic-based interventions provide a politically neutral path forward. Unlike fact-checks that require adjudicating contested claims, logic-checks focus on argumentative structure—a dimension people can evaluate across belief divides. Our logic warnings improved discrimination uniformly for Democrats and Republicans. This political neutrality makes logic-based interventions particularly promising for polarized contexts and organisations resistant to politically-aligned interventions. Given that most anti-climate arguments contain argumentative fallacies [36,38], logic-based interventions present a particularly promising method for decreasing the spread of climate misinformation.

Second, logic-based interventions are highly scalable. Emerging work shows that AI systems can detect fallacies in climate-related arguments, pointing to a path that avoids the high costs and bias accusations associated with centralized fact-checking. Moreover, our results suggest that even minimal, passive engagement with logical warnings improved discernment (by about 5%). While only a few participants sought additional information (5%), those who did showed even greater improvements, around 21%. This suggests that embedding explanations directly—making them unavoidable rather than optional—could optimize effectiveness. Whether through AI detection systems or community notes-style crowdsourced identification, highlighting logical quality presents a promising, evidence-based method for further investigation.

This set of results presents a promising picture, suggesting that improving climate argumentation does not require overcoming intractable identity-based biases. People retain the capacity to distinguish strong from weak arguments even in polarized domains—they simply need contexts that direct attention to argument structure. Our demonstration that minimal interventions can shift reasoning across political divides suggests that addressing climate misbeliefs at scale is achievable through context design rather than individual persuasion.

Methods

Study 1

Participants

In total, 1631 participants started the experiment and gave consent, 115 could not participate due to giving the incorrect attention check answer (and hence produced no further data). Finally, due to a bug in the experiment, people with Safari browser could not access the experiments, so 320 dropped out, hence, altogether, we collected data from 1196 participants (639 female, 545 male and 8 identified as transgender or non-binary; Mean age = 48.4 years, SD = 17.1 years). Participants were recruited with Lucid, aiming for national representation for the US (quota-matched for age, gender, ethnicity). They were paid $1.5 for participation.

Materials

Reasoning problems. In total, we created 96 reasoning problems; out of which 48 were about climate change and 48 about less-politicised science or technology topics, namely, risks posed by AI, GMO and nanotechnology. For climate change arguments, we manipulated 2 aspects: 1) whether they were pro or contra climate change[1] (directionality) and 2) whether they were strong or weak (strength). **Directionality manipulation:** Each problem argued for or against one aspect of climate change: that is, that climate change is happening, that it is caused by human activity or that it is a risk. We had eight arguments pro and eight arguments contra for each three aspects. **Strength manipulation:** We manipulated the strength of the arguments by introducing four types of fallacies: denying the antecedent, affirming the consequent, post hoc ergo propter hoc, and overgeneralization/oversimplification. We will refer to arguments with the fallacies as weak arguments and arguments without fallacies as strong, solely referring to the logical fallacies in them. All problems can be found on GitHub. Altogether, we have 4 arguments in each category: 4 strong, arguing that that climate change is happening (pro), 4 strong arguing that climate change is not happening (contra) and so on.. From each of these categories, participants were randomly presented to 2 problems. Similarly, we applied the same manipulations to the non-politized arguments, whereas these statements argued for or against the riskiness of AI, riskiness of GMO, or the riskiness of nanotechnology. With regard to each argument, participants had to answer two questions. First, 'How strong is the argument?',clicking on either Weak or Strong. Then, participants were asked 'How difficult was it to make this judgment?, selecting from a scale of1 - very easy to 7 - very difficult '.

Prior beliefs. We measured prior beliefs with 3 different questions according to the argument contents. As arguments were threefold: 1) climate change is happening, 2) climate change is

[1] Pro climate change arguments affirm the idea that climate change is happening, anthropogenic, or risky, and vice-versa for contra climate change arguments.

caused by human activity, 3) climate change is a threat, we used matching prior belief questions specifically asking about these aspects. Participants had to indicate their opinion on a scale from 0 to 100 on the following questions:

1. Climate change is happening: *'Global warming refers to the idea that the world's average temperature has been increasing over the past 150 years, may be increasing more in the future, and that the world's climate may change as a result. What do you think, what is the probability that global warming is happening?' [0: Definitely NOT happening, 50: Neutral/equally likely, 100: Definitely happening]*
2. Climate change is caused by human activity: *'What is the probability that human activity is the main cause behind global warming?' [0: Definitely NOT the main cause, 50: Neutral/equally likely, 100: Definitely the main cause]*
3. Climate change is a threat: *'How significant is the threat climate change poses to...'*
   ... human safety [0: Very insignificant, 100: Very significant]
   ...global biodiversity/ecosystems [0: Very insignificant, 100: Very significant]
   ...agricultural productivity [0: Very insignificant, 100: Very significant]
   ...world and national economies [0: Very insignificant, 100: Very significant]

This latter question asked about beliefs regarding aspects of climate-change threat as the reasoning problems were also specifically mentioning these threats.

Importantly, for the statistical analysis, we created one prior belief value for each trial depending on the contents of the argument. That is, if the content of the argument argued for or against the idea that climate change poses a threat to agricultural productivity, for that specific trial, the prior belief value was the participant's response to the corresponding prior belief question about climate change posinga significant threat to agricultural productivity.

Political partisanship. We measured political partisanship with a simple question taken from Bago et al., 2023): 'Which of the following best describes your political preference?' with answer options: Strong Republican, Republican, Lean Republican, Lean Democrat, Democrat, Strong Democrat.

Other measures. We asked participants about basic demographics, namely: gender, age, highest level of education, as well as their economic and social ideology. Participants also had to fill out a 7-item version of the Cognitive Reflection Test (CRT).

Procedure

After giving consent to participate in the experiment, participants were asked a simple attention check question: 'Puppy is to dog as kitten is to …' ;if they did not give the correct solution (cat), they could not participate in the experiment. Then, participants were asked demographic questions, prior belief, partisanship, and ideology questions, and finally, the CRT. After these, participants were presented with the following instructions before they started with the reasoning problems:

'In this experiment, you will see 48 statements about science and technology. We want to understand how strong you think those statements or arguments are. After you read the statement, you will have to click on either "Strong" or "Weak" and then the experiment will automatically advance to the next statement. We will also ask you how difficult it was for you to make this decision after each statement. Click on "Next" to start the experiment!'

Participants were then presented with the selected statements in a randomised order, one at a time.

Statistical analysis

We used logistic mixed effect models to analyze the data and applied maximal random effect structure. That is, we added the random intercept of subjects and items (each item had a unique identifier), and we allowed the intercept of subjects to vary over argument strength and argument direction, while allowing the random intercept of items to vary over partisanship and prior beliefs. This approach accounts for individual and item-level variability in responses, improving generalizability across items and reducing the risk of inflated Type I error rates. We coded argument strength as 0.5 for strong and -0.5 for weak arguments, direction as -0.5 if argument is pro and 0.5 if argument is contra. Prior belief was centered on 0 on a scale from -1 (climate skeptic) to +1 (climate believer). Partisanship was also centered on 0 and values were transformed to a scale from -1 (Strong Republicans) to +1 (Strong Democrats). All analyses reported here were preregistered.

Study 2 and 3

Participants

In Study 2, 1832 participants started the experiment and gave consent, while 147 could not participate due to giving the incorrect attention check answer (and hence produced no data). Finally, due to a coding mistake in the experiment, people with Safari browser could not access the experiments, so they dropped out; 758 dropped out, hence, altogether, we collected data from 927 participants (419 female, 492 male and 9 identified as transgender or non-binary, Mean age = 47.6 years, SD = 16.1 years). Participants were recruited with Lucid, and are representative for US population (quota-matched for age, gender, ethnicity). In total, 469 participants took part in the control group and 452 in the treatment condition (identity threat manipulation). They were paid $1.5 for participation.

In Study 3, 1599 participants started the experiment and gave consent, while 189 could not participate due to giving the incorrect attention check answer (and hence produced no data).

Altogether, we collected data from 1060 participants as 350 left before the experiment started (644 female, 413 male and 3 identified as transgender or non-binary, Mean age = 52.3 years, SD = 16.7 years). Participants were recruited with Lucid, and are representative for US population (quota-matched for age, gender, ethnicity). In total, 546 participants took part in the control group and 514 in the treatment condition (logical cues manipulation).They were paid $1.5 for participation.

Materials & Procedure

Reasoning problems. Participants were presented with the same set of reasoning problems as in Study 1. As these studies were longer (and studies on Lucid cannot take more than 10 minutes), we decreased the amount of arguments people got from the non-politicized category; they only saw 12 non-politized items and 24 climate change items.

Identity threat (Study 2). Participants were randomly assigned between-subjects to the control (same as Study 1) or the identity threat condition. In the partisan cues condition participants were always presented with the target argument and a partisan cue argument, as if it were a quote retweet text, above the target argument (see Figure 1 for examples). To assure participants actually read this retweet text before the target argument, they were presented with this retweet for 5 seconds before the questions and the target argument appeared. The retweet always included an ad hominem attack toward either Republicans or Democrats, but provided no logical[2] counterargument to the content of the target argument.. To increase ecological validity and ensure identity threat was activated, we implemented a manipulation that ad hominem attacks their political identity in an ecologically valid context (retweets). This manipulation was designed while bearing in mind the existing manipulations in the partisanship-motivated reasoning literature. Therein, they typically instruct participants to view the following information through political lenses, thus prompting them for political identity protection (Taber & Lodge, 2006). Such instructions are not ecologically valid, as in real social media environments, there are no direct instructions for political thinking. Indeed, by definition, identity-protection should be triggered when arguments make an ad hominem attack against one's identity. Hence, we created a manipulation that attacks people's political identity directly with ad hominem arguments, while also being externally valid and taking the form of quote retweets. The full list of argument-retweet matching can be found on GitHub. These hypothetical retweets always attacked the hypothetical person making the original arguments, for example:

Retweet: *'Typical Democrats, using natural disasters to push their climate change hoax. They never let a crisis go to waste.'*

Target argument: *'Climate change is real, otherwise we would not be seeing all of these natural disasters on the news.'*

[2] That is, the retweet did not contain any information that would explicitly help the participant to identify a logical flaw in the reasoning of the tweet, if there was one.

Hence, we always created retweets that attacked Republicans for anti-climate arguments and retweets that attacked Democrats for pro-climate arguments.

Before the experiment, we presented people with the following instructions in the partisan cues condition:

'In this experiment, you will see 36 statements (in the form of tweets) about science and technology. These statements are accompanied by comments from people who re-shared them - just like you would see on X (formerly Twitter) or other social media such as Facebook. We want to know what you think of these tweets in the light of the accompanying comment.

This is how the trials will look: [picture of trial]

We want to understand how strong you think the original statements or arguments are. You will have to click on either "Weak" or "Strong". We will also ask you how difficult it was for you to make this decision after each statement. Once you have made up your mind about strength and difficulty, you can click on Next Tweet.

To ensure that you read both texts, we will first show you the comment for 5 seconds, and then the comment and the questions after that.

Click on "Next" to start the experiment!'

People in the control condition received the same instructions as in Study 1.

Logical warnings (Study 3). Participants were randomly assigned between-subjects to the control (same as Study 1) or the logical cues condition. In the logical cues condition, participants were informed that they may receive AI-generated warnings, then when participants were presented with an argument containing an argumentative fallacy, they also received a warning: '*Warning: This tweet may contain a logical fallacy. Click here for more details.*' Participants could also click on the button to learn more if they were interested. This pop-up then contained a content specific explanation of the fallacy. The warnings contained a clickable link over a direct description of the content-based fallacy refutation to mimic more closely how social media implements such warnings, like fact-checking information, whereas a warning is placed, but the detailed fact-checking arguments can be reviewed upon clicking on it. Arguments with no argumentative fallacies did not get any warnings, reflecting the fact that even though they might not contain argumentative fallacies, they could still be false (just non-checked) or rest on false premises, avoiding that participants do not perceive these arguments as strong all the time, similarly to experiments testing the effectiveness of fact-checking labels on news discernment (Martel & Rand, 2024). As conveyed, these warnings were generated byGPT-4o. In addition to adding ecological validity, this design allowed for additional tests to rule out potential demand effects, as under a demand effect, one could expect no difference in strength judgments between trials when the pop-up was clicked, versus was not clicked, as strength judgments would be at the minimum already.

In the logical cues treatment, participants were warned about the logical fallacies and received the following instructions beforehand:

*'In this experiment, you will see 36 statements (in the form of tweets) about science and technology. <u>Some</u> of these statements may contain logical fallacies. A **fallacy** is the use of invalid or otherwise faulty reasoning in the construction of an argument.*

*For example, the statement "Aliens must exist because no one has proven they don't" commits the **Appeal to Ignorance** fallacy. This is faulty reasoning because a lack of evidence against something does not prove it to be true, just as the inability to disprove unicorns doesn't mean they exist.*

<u>Some</u> of the statements were checked by an Artificial Intelligence for such argumentative fallacies. When the AI found a fallacy, it created a short warning and a clickable link where the fallacy is explained. If you want to know more, you can click on the link. However, it is not compulsory in this experiment.

This is how a trial with a fallacy warning will look: [picture of a trial]

We want to understand how strong you think the statements or arguments are. We are interested in your true opinion about the statement, NOT the warning. You will have to click on either "Weak" or "Strong". We will also ask you how difficult it was for you to make this decision after each statement. Once you have made up your mind about strength and difficulty, you can click on Next Tweet.

Click on "Next" to start the experiment!'

To ensure that effects cannot simply be driven by the instructions in the logical cues condition (that people were told about logical fallacies), we also made the same warning in the control condition instructions:

*'In this experiment, you will see 36 statements (in the form of tweets) about science and technology. <u>Some</u> of these statements may contain logical fallacies. A **fallacy** is the use of invalid or otherwise faulty reasoning in the construction of an argument.*

*For example, the statement "Aliens must exist because no one has proven they don't" commits the **Appeal to Ignorance** fallacy. This is faulty reasoning because a lack of evidence against something does not prove it to be true, just as the inability to disprove unicorns doesn't mean they exist.*

We want to understand how strong you think the original statements or arguments are. You will have to click on either "Weak" or "Strong". We will also ask you how difficult it was for you to make this decision after each statement. Once you have made up your mind about strength and difficulty, you can click on Next Tweet.

Click on "Next" to start the experiment!'
Other measures. The same demographic, partisanship and prior measures were applied as in Study 1. Study 2 did not contain the Cognitive Reflection Test in order to make the experiment shorter to fit the 10-minute Lucid criteria.

Statistical analysis

Statistical analysis followed the same scheme as for Study 1. We used logistic mixed effect models to analyze the data and applied maximal random effect structure. That is, we added the random intercept of subjects and items (each item had a unique identifier), and we allowed the intercept of subjects to vary over argument strength and argument direction, while allowing the random intercept of items to vary over partisanship and prior beliefs. We coded argument strength as 0.5 for strong and -0.5 for weak arguments, direction as -0.5 if argument is pro and 0.5 if argument is contra. Prior belief was centered on 0 on a scale from -1 (climate skeptic) to +1 (climate believer). Partisanship was also centered on 0 and values were transformed to a scale from -1 (Strong Republicans) to +1 (Strong Democrats). Treatment was dummy coded 0 representing the control condition and 1 as the treatment (partisan cues in Study 2 and logical cues in Study 1). All analyses reported here were preregistered.

Code, data and materials availability

Code, data and all materials are available on GitHub:
https://github.com/bencebago/climate_argumentation

Preregistration

Preregistration for Study 1 available at: https://aspredicted.org/449q-wh8y.pdf Preregistration for Study 2 available at: https://aspredicted.org/vw52nr.pdf Preregistration for Study 3 available at: https://aspredicted.org/cq9468.pdf

Note that we preregistered analysis on difficulty scores, which are not reported here, as they will be analyzed within another project.

References

1. Rojas, C. *et al.* Hierarchical machine learning models can identify stimuli of climate change misinformation on social media. *Commun. Earth Environ.* **5**, 436 (2024).

2. Falkenberg, M. *et al.* Growing polarization around climate change on social media. *Nat. Clim. Change* **12**, 1114–1121 (2022).

3. Nwokolo, S. C. Climate hoax: The shift from scientific discourse to speculative rhetoric in climate change conversations. *Res.* **2**, 100322 (2025).

4. Nicolosi, E., Medina, R., Brewer, S., Vorkink, M. & Allred, A. The new denial: climate solution misinformation on social media. *Glob. Sustain.* **8**, e31 (2025).

5. Aspernäs, J., Erlandsson, A. & Nilsson, A. Motivated formal reasoning: Ideological belief bias in syllogistic reasoning across diverse political issues. *Think. Reason.* **29**, 43–69 (2023).

6. Bohr, J. Public views on the dangers and importance of climate change: predicting climate change beliefs in the United States through income moderated by party identification. *Clim. Change* **126**, 217–227 (2014).

7. Caddick, Z. A. & and Feist, G. J. When beliefs and evidence collide: psychological and ideological predictors of motivated reasoning about climate change. *Think. Reason.* **28**, 428–464 (2022).

8. Calvillo, D. P., Swan, A. B. & Rutchick, A. M. Ideological belief bias with political syllogisms. *Think. Reason.* **26**, 291–310 (2020).

9. Fielding, K. S. & Hornsey, M. J. A Social Identity Analysis of Climate Change and Environmental Attitudes and Behaviors: Insights and Opportunities. *Front. Psychol.* **7**, (2016).

10. Gampa, A., Wojcik, S. P., Motyl, M., Nosek, B. A. & Ditto, P. H. (Ideo)Logical Reasoning: Ideology Impairs Sound Reasoning. *Soc. Psychol. Personal. Sci.* **10**, 1075–1083 (2019).

11. Guilbeault, D., Becker, J. & Centola, D. Social learning and partisan bias in the interpretation of climate trends. *Proc. Natl. Acad. Sci.* **115**, 9714–9719 (2018).

12. Kahan, D. M. The politically motivated reasoning paradigm, Part 2: Unanswered questions. *Emerg. Trends Soc. Behav. Sci. Interdiscip. Searchable Linkable Resour.* 1–15 (2015).

13. Kahan, D. M. *et al.* The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nat. Clim. Change* **2**, 732–735 (2012).

14. Keller, L., Hazelaar, F., Gollwitzer, P. M. & Oettingen, G. Political ideology and environmentalism impair logical reasoning. *Think. Reason.* **30**, 79–108 (2024).

15. Rutjens, B. T., Sutton, R. M. & Van Der Lee, R. Not All Skepticism Is Equal: Exploring the Ideological Antecedents of Science Acceptance and Rejection. *Pers. Soc. Psychol. Bull.* **44**, 384–405 (2018).

16. Ucar, G. K., Yalcin, M. G., Planalı, G. Ö. & Reese, G. Social identities, climate change denial, and efficacy beliefs as predictors of pro-environmental engagements. *J. Environ. Psychol.* **91**, 102144 (2023).

17. Bago, B., Rand, D. G. & Pennycook, G. Reasoning about climate change. *PNAS Nexus* **2**, pgad100 (2023).

18. Cook, J. & Lewandowsky, S. Rational Irrationality: Modeling Climate Change Belief Polarization Using Bayesian Networks. *Top. Cogn. Sci.* **8**, 160–179 (2016).

19. Deans-Browne, C. & Singmann, H. For everyday arguments prior beliefs play a larger role on perceived argument quality than argument quality itself. *Cognition* **266**, 106257 (2026).

20. Gerber, A. & Green, D. MISPERCEPTIONS ABOUT PERCEPTUAL BIAS. *Annu. Rev. Polit. Sci.* **2**, 189–210 (1999).

21. Hahn, U. Argument Quality in Real World Argumentation. *Trends Cogn. Sci.* **24**, 363–374 (2020).

22. Hahn, U. & Harris, A. J. L. What Does It Mean to be Biased. in *Psychology of Learning and Motivation* vol. 61 41–102 (Elsevier, 2014).

23. Hahn, U. & Oaksford, M. The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychol. Rev.* **114**, 704–732 (2007).

24. Mercier, H. The Argumentative Theory: Predictions and Empirical Evidence. *Trends Cogn. Sci.* **20**, 689–700 (2016).

25. Tappin, B. M., Pennycook, G. & Rand, D. G. Rethinking the link between cognitive sophistication and politically motivated reasoning. *J. Exp. Psychol. Gen.* **150**, 1095–1114 (2021).

26. De Neys, W. Advancing theorizing about fast-and-slow thinking. *Behav. Brain Sci.* 1–68 (2022) doi:10.1017/S0140525X2200142X.

27. Evans, J. St. B. T. & Stanovich, K. E. Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspect. Psychol. Sci.* **8**, 223–241 (2013).

28. Johnson-Laird, P. N., Goodwin, G. P. & Khemlani, S. S. Mental models and reasoning. in *International Handbook of Thinking and Reasoning* (Routledge, 2017).

29. Thompson, V. A., Prowse Turner, J. A. & Pennycook, G. Intuition, reason, and metacognition. *Cognit. Psychol.* **63**, 107–140 (2011).

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

31. Purcell, Z. A., Roberts, A. J., Handley, S. J. & Howarth, S. Eye Movements, Pupil Dilation, and Conflict Detection in Reasoning: Exploring the Evidence for Intuitive Logic. *Cogn. Sci.* **47**, e13293 (2023).

32. Trippas, D., Handley, S. J., Verde, M. F. & Morsanyi, K. Logic brightens my day: Evidence for implicit sensitivity to logical validity. *J. Exp. Psychol. Learn. Mem. Cogn.* **42**, 1448–1457 (2016).

33. Boissin, E., Caparos, S., Abi Hana, J., Bernard, C. & De Neys, W. Easy-fix attentional focus manipulation boosts the intuitive and deliberate use of base-rate information. *Mem. Cognit.* **53**, 995–1007 (2025).

34. Purcell, Z. A., Wastell, C. A. & Sweller, N. Domain-specific experience and dual-process thinking. *Think. Reason.* **27**, 239–267 (2021).

35. Purcell, Z. A., Wastell, C. A. & Sweller, N. Eye movements reveal that low confidence precedes deliberation. *Q. J. Exp. Psychol.* **76**, 1539–1546 (2022).

36. Cook, J., Ellerton, P. & Kinkead, D. Deconstructing climate misinformation to identify reasoning errors. *Environ. Res. Lett.* **13**, 024018 (2018).

37. Lewandowsky, S. Climate Change Disinformation and How to Combat It. *Annu. Rev. Public Health* **42**, 1–21 (2021).

38. Zanartu, F., Cook, J., Wagner, M. & García, J. A technocognitive approach to detecting fallacies in climate misinformation. *Sci. Rep.* **14**, 27647 (2024).

39. Pennycook, G., Bago, B. & McPhetres, J. Science beliefs, political ideology, and cognitive sophistication. *J. Exp. Psychol. Gen.* **152**, 80–97 (2023).

40. Taber, C. S. & Lodge, M. Motivated Skepticism in the Evaluation of Political Beliefs. *Am. J. Polit. Sci.* **50**, 755–769 (2006).

41. Rains, S. A., Kenski, K., Coe, K. & Harwood, J. Incivility and Political Identity on the Internet: Intergroup Factors as Predictors of Incivility in Discussions of News Online. *J. Comput.-Mediat. Commun.* **22**, 163–178 (2017).

42. Suhay, E., Bello-Pardo, E. & Maurer, B. The Polarizing Effects of Online Partisan Criticism: Evidence from Two Experiments. *Int. J. Press.* **23**, 95–115 (2018).

43. Bavel, J. J. V. & Pereira, A. The Partisan Brain: An Identity-Based Model of Political Belief. *Trends Cogn. Sci.* **22**, 213–224 (2018).

44. Kahan, D. M. Ideology, motivated reasoning, and cognitive reflection. *Judgm. Decis. Mak.* **8**, 407–424 (2013).

45. Slothuus, R. & de Vreese, C. H. Political Parties, Motivated Reasoning, and Issue Framing Effects. *J. Polit.* **72**, 630–645 (2010).

46. Martel, C. & Rand, D. G. Fact-checker warning labels are effective even for those who distrust fact-checkers. *Nat. Hum. Behav.* **8**, 1957–1967 (2024).

47. Porter, E. & Wood, T. J. The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proc. Natl. Acad. Sci.* **118**, e2104235118 (2021).

48. Walter, N., Cohen, J., Holbert, R. L. & Morag, Y. Fact-Checking: A Meta-Analysis of What Works and for Whom. *Polit. Commun.* **37**, 350–375 (2020).

49. Mena, P. Reducing Misinformation Credibility: How Explanations Impact the Effectiveness of Social Media Warning Labels and Fact-Checking Source Recall. *Journal. Mass Commun. Q.* 10776990251347657 (2025) doi:10.1177/10776990251347657.

50. Ackerman, R. & Thompson, V. A. Meta-Reasoning: Monitoring and Control of Thinking and Reasoning. *Trends Cogn. Sci.* **21**, 607–617 (2017).

51. De Neys, W. *Dual Process Theory 2.0*. (Routledge, London, 2018). doi:10.4324/9781315204550.

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

53. Handley, S. J. & Trippas, D. Dual Processes and the Interplay between Knowledge and Structure: A New Parallel Processing Model. in *Psychology of Learning and Motivation* (ed. Ross, B. H.) vol. 62 33–58 (Academic Press, 2015).