

Simple logic warnings improve climate argument evaluation across beliefs and partisanship

Received: 8 December 2025

Accepted: 19 August 2026

Published online: 21 September 2026

 Check for updates

Bence Bago ¹ & Zoe A. Purcell ^{2,3}

Collective climate action requires shared beliefs aligned with scientific consensus, yet fallacious information leads people to misjudge arguments. The mechanisms behind these failures remain unclear. Here we isolate the independent effects of partisan identity, prior beliefs and cognitive limits on argument-quality detection. Across four preregistered studies ($N = 4,000$), we introduce a protocol that orthogonally manipulates argument quality and direction across realistic contexts. In neutral contexts, prior beliefs dominated judgements, with partisanship and argument quality having smaller effects. Contrary to identity-protective accounts, embedding arguments in hostile partisan contexts did not increase partisan bias or impair argument-quality detection. Instead, partisan attacks reduced reliance on prior beliefs, prompting participants to move away from their prior positions. Finally, minimal logic warnings indicating possible fallacies enhanced discrimination between weak and strong arguments. These findings show climate argument evaluation is context sensitive and can be improved with simple logic warnings.

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 low-quality (for example, logically flawed, argumentatively erroneous) anti-climate arguments that continue to gain traction despite their evident weaknesses^{1–4}. Understanding why people fail to discard such arguments and whether their evaluations can be improved requires identifying the key factors that influence argumentation and how those factors can be leveraged to improve climate argument evaluation.

Two camps of research have sought to explain how people evaluate climate arguments and why they often err. ‘Identity-protection’ theories suggest that people are motivated to align their beliefs with those of their partisan group—argumentation is influenced by argument-identity alignment^{5–18}. Climate change has become a politically polarized issue; in the USA, Republicans are substantially less likely than Democrats to accept the scientific consensus. Studies have shown that partisanship predicts climate attitudes beyond general

knowledge or reasoning ability, and that partisan framing increases motivated evaluation^{5,10,11,13,15}. Thus, identity-protective theories suggest identity alignment is the primary driver of climate argumentation errors^{7,9,12,13,15}. Conversely, ‘coherence-based’ theories suggest that people are resistant to changing their minds—argument evaluation is influenced by prior-belief alignment^{19–26}. Studies show that people rate prior-consistent arguments as stronger than belief-inconsistent arguments across a range of political topics, regardless of argument quality²¹. Hence, coherence-based theories purport that prior beliefs can corrupt argumentation. Both identity-protection and coherence-based theories minimize the role of argument quality.

However, studies of reasoning show that people are sensitive to argument quality^{27–30}. Argument quality is the degree to which an argument’s reasoning supports its conclusion²³ and is commonly operationalized by including or excluding formal or informal fallacies. Fallacies are patterns that weaken the inferential link between premises and conclusion³¹. It is not clear, however, how argument quality interacts

¹Department of Social Psychology, Tilburg University, Tilburg, the Netherlands. ²LaPsyDÉ, Université Paris Cité, CNRS, Paris, France. ³Department of Psychology, Health and Gender, The American University of Paris, Paris, France. ✉ e-mail: b.bago@tilburguniversity.edu

with prior beliefs and partisanship, and whether it can be harnessed to improve climate argumentation.

To answer these questions, three gaps in the literature must be addressed. First, the relative causal contribution of priors and partisanship are unknown. These factors are rarely considered together and most designs are correlational^{19,26} (cf. ref. 32). Second, the independent role of argument quality and its interaction with priors and partisanship is unknown, despite anti-climate arguments disproportionately relying on identifiable fallacies^{33–35}. Third, whether this natural co-occurrence of low argument quality and anti-climate content can be utilized for intervention remains untested.

Addressing these gaps requires overcoming several challenges. First, for climate arguments, partisanship and prior beliefs naturally covary. In the USA, Republicans are less likely than Democrats to accept climate science, confounding identity with priors. Furthermore, anti-climate arguments contain fallacies, such as false dichotomies, circular reasoning or emotionally charged appeals, more often than pro-climate arguments³⁵, creating a confound between argument direction and quality. Consequently, existing studies have not determined the independent, causal contributions of these factors.

Second, causal, single-factor tests use unrealistic manipulations and inappropriate measures. Manipulations to trigger identity-protective reasoning are highly contrived and unlikely to generalize to naturalistic settings (for example, instructing participants to consider the political landscape before responding)³⁶. Moreover, such manipulations may not engage the intended psychological processes: viewing issues politically does not necessarily threaten one's identity. Manipulations to improve rational reasoning use abstract tasks with normative probabilistic or formal rules, and effectiveness on argumentation is inconsistent^{27,37}. Prior beliefs are weakly matched with arguments (for example, aggregated climate belief scales are used as priors across specific arguments such as whether climate change is a threat¹⁹). Thus, while partisanship, priors and argument quality may each influence climate argumentation, their relative, causal roles in realistic settings remain unknown.

Paradigm

To overcome these challenges, we developed a protocol that tightly controls each factor within ecologically motivated experiments—that is, within experiments and manipulations that are designed to emulate realistic contexts while maintaining experimental control. Participants first report their partisan identity and their topic-specific climate beliefs (about climate change's existence, anthropogenic cause and potential threat). They then evaluate arguments that are fully crossed in direction (pro- or anti-climate), topic (existence, cause or threat) and quality (with or without a fallacy). Because priors are measured at the topic level, the relative effects of partisanship and argument quality cannot be attributed to a mismatch between prior and argument topic.

Four experiments (total $N = 4,000$) used this protocol. Study 1 established the baseline contributions of identity, priors and argument quality in a neutral context without additional partisan or logic cues. Motivated by politicized social media discourse, Studies 2a and 2b tested whether identity threat increased reliance on partisanship, by accompanying arguments with contesting (2a) or endorsing (2b) retweets. Motivated by monitored social discourse, Study 3 tested whether increasing logic salience improved argument evaluation, by warning participants that a fallacy might be present. Samples were nationally representative for age, gender and ethnicity (see Fig. 1 for example items and Methods for full details).

These designs allow us to assess the relative contributions of priors, partisanship and argument quality to argument evaluation. Because the influence of priors and partisanship depends on argument direction, their interaction with direction indicates that belief- or identity-consistent arguments are rated strong more often than inconsistent ones. Therefore, we focus on three effects: prior $\times$ direction

($pr \times d$), partisanship $\times$ direction ($pa \times d$) and the main effect of argument quality (a), which does not depend on direction. In Studies 2a, 2b and 3, we examine how each is modulated by the identity-threat and logic-salience manipulations—for example, treatment $\times$ ($pr \times d$) captures how identity threat shifts reliance on prior beliefs. Full models and higher-order effects are reported in Supplementary Information, though these should be interpreted cautiously, as the studies were not designed or necessarily powered for them.

Prior beliefs dominate climate argument evaluation in a neutral context

In Study 1, consistent with coherence-based accounts, priors had the largest effect: belief-consistent arguments were rated as strong more often than belief-inconsistent ones ($pr \times d: \beta = 2.76, P < 0.001$). Partisanship showed a smaller but reliable effect: partisan-aligned arguments were rated as strong more often than misaligned arguments ($pa \times d: \beta = 0.79, P < 0.001$). The effect of argument quality was smaller but significant: strong arguments were rated as strong more often than weak ones ($a: \beta = 0.41, P = 0.006$) (Fig. 2).

We then examined whether the effect of argument quality varied with priors or politics. Argument quality and priors interacted significantly ($\beta = 0.73, P < 0.001; a \times pr \times d$): people were better at discerning weak from strong arguments for prior-aligned arguments ($\beta = 0.65, P < 0.001$) than for misaligned arguments, though discernment remained significant even then ($\beta = 0.36, P = 0.014$). In contrast, argument quality did not interact with partisanship, $\beta = -0.12, P = 0.370$ ($a \times pa \times d$) (Extended Data Fig. 1, Supplementary Tables 2–5 and Supplementary Figs. 5 and 6).

Identity threat shifts judgements, but not towards partisan entrenchment

To test the causal role of partisanship on argument evaluation, we maximized identity threat via contesting or endorsing retweets. Across two studies, participants were randomly assigned to view climate arguments alongside hostile retweets (treatment) arguments in isolation (control). In Study 2a, the retweet contested the target argument while denigrating the associated party, attacking Democrats alongside pro-climate arguments and Republicans alongside anti-climate arguments (Fig. 1b, examples A and B). This quote-share format mirrors social media contexts where climate arguments are frequently embedded in partisan rhetoric; precisely the contexts where identity-protective reasoning is theorized to be most influential^{2,11,38,39}. Because partisanship and climate belief are correlated, every attack in Study 2a fell on a belief-consistent argument. Therefore, Study 2b reversed this pairing: the retweet endorsed the target argument while denigrating the opposing party, such that it was attacking Democrats with anti-climate arguments and Republicans with pro-climate arguments (Fig. 1b, examples C and D). Hence, these designs offer greater ecological relevance and avoid previous confounds.

Partisanship could impact evaluation directly—party-aligned arguments are rated as strong more often than unaligned arguments, regardless of their prior beliefs^{13,17}—or indirectly—reliance on prior beliefs increases for party-aligned arguments¹⁸. Both pathways converge on one prediction: under identity threat, people should reinforce their existing positions, rating identity- or belief-consistent arguments as strong and inconsistent ones as weak more often than in neutral contexts.

As in Study 1, further challenging identity-protective accounts, priors had the largest effect on argument evaluation ($pr \times d$: Study 2a, $\beta = 2.85, P < 0.001$; Study 2b, $\beta = 3.16, P < 0.001$), followed by partisan identity ($pa \times d$: Study 2a, $\beta = 0.73, P < 0.001$; Study 2b, $\beta = 0.76, P < 0.001$) and a smaller effect of argument quality (a : Study 2a, $\beta = 0.33, P = 0.047$; Study 2b, $\beta = 1.01, P < 0.001$; see Fig. 2 for standardized direct effects). Notably, the argument-quality effect was larger in Study 2b than Study 2a, possibly because Prolific participants in Study 2b paid

greater attention than the Lucid participants in Study 2a, as reflected in their lower attention-check failure rate (Methods).

Strikingly, contrary to predictions of the identity-protective account, the effect of prior beliefs shifted under threat in both studies in opposite directions, but always away from participants' starting positions (treatment $\times$ (pr $\times$ d): Study 2a, $\beta = -1.03$, $P < 0.001$; Study 2b, $\beta = -0.60$, $P < 0.001$; Fig. 3). In Study 2a, where retweets contested arguments, participants downgraded belief-consistent arguments (simple effect, $\beta = -0.54$, $P < 0.001$) while leaving belief-inconsistent arguments unchanged (simple effect, $\beta = -0.01$, $P = 0.912$). In Study 2b, where retweets endorsed arguments, participants upgraded belief-inconsistent arguments (simple effect, $\beta = 0.22$, $P = 0.046$) while leaving belief-consistent ones unchanged (simple effect, $\beta = -0.10$, $P = 0.201$). In neither study did the effect of argument-quality discernment change under threat (treatment $\times$ a: Study 2a, $\beta = -0.01$, $P = 0.921$; Study 2b, $\beta = -0.05$, $P = 0.663$). Partisanship behaved differently across the two studies: it did not significantly shift under threat in Study 2a (treatment $\times$ (pa $\times$ d): $\beta = 0.02$, $P = 0.913$) but it did in Study 2b (treatment $\times$ (pa $\times$ d): $\beta = 0.68$, $P < 0.001$). Post hoc simple-effects analyses (Supplementary Table 26) showed that neither party systematically strengthened its own identity-congenial arguments under threat. In Study 2a, treatment lowered perceived strength across the board, significantly so only for climate-affirming arguments for Democrats and Republicans alike. In Study 2b, treatment raised perceived strength of climate-sceptical arguments (significantly only for Democrats), while climate-affirming arguments fell for Republicans and were essentially unchanged for Democrats (Supplementary Information). In short, the shifts ran counter to identity-protective predictions and did not consistently widen partisan differences. Overall, these patterns do not match identity-protective predictions for either direct or indirect causal pathways; the shifts observed under threat moved participants away from their prior-consistent positions or towards the endorsed argument rather than towards the participant's own party.

From an identity-protective perspective, these results are unexpected. We speculate that the social signal in the retweets—whether others contest or endorse a position—may drive these effects, and not identity. Contesting retweets increased the number of strong judgements (Study 2a) and endorsing ones reduced them (Study 2b). The non-climate arguments let us probe this further. For participants in the treatment condition, less-politicized items were accompanied by contesting comments that did not mention politics (for example, for genetically modified organisms: 'Indeed! Why label them if they're so safe? The industry crowd hates labels because they expose what corporate agriculture is hiding. Shameless people.'). Despite the lack of partisan threat, similar trends emerged. In Study 2a, contesting comments reduced reliance on priors (treatment $\times$ (pa $\times$ d): $\beta = -0.43$, $P = 0.056$) without a uniform change in perceived strength

(main effect of treatment: $\beta = -0.11$, $P = 0.218$; Supplementary Table 8); in Study 2b, endorsing comments produced a uniform increase in perceived strength (main effect of treatment: $\beta = 0.20$, $P = 0.002$) with no shift in reliance on priors (treatment $\times$ (pr $\times$ d): $\beta = 0.04$, $P = 0.818$; Supplementary Table 12). However, this post hoc analysis should be

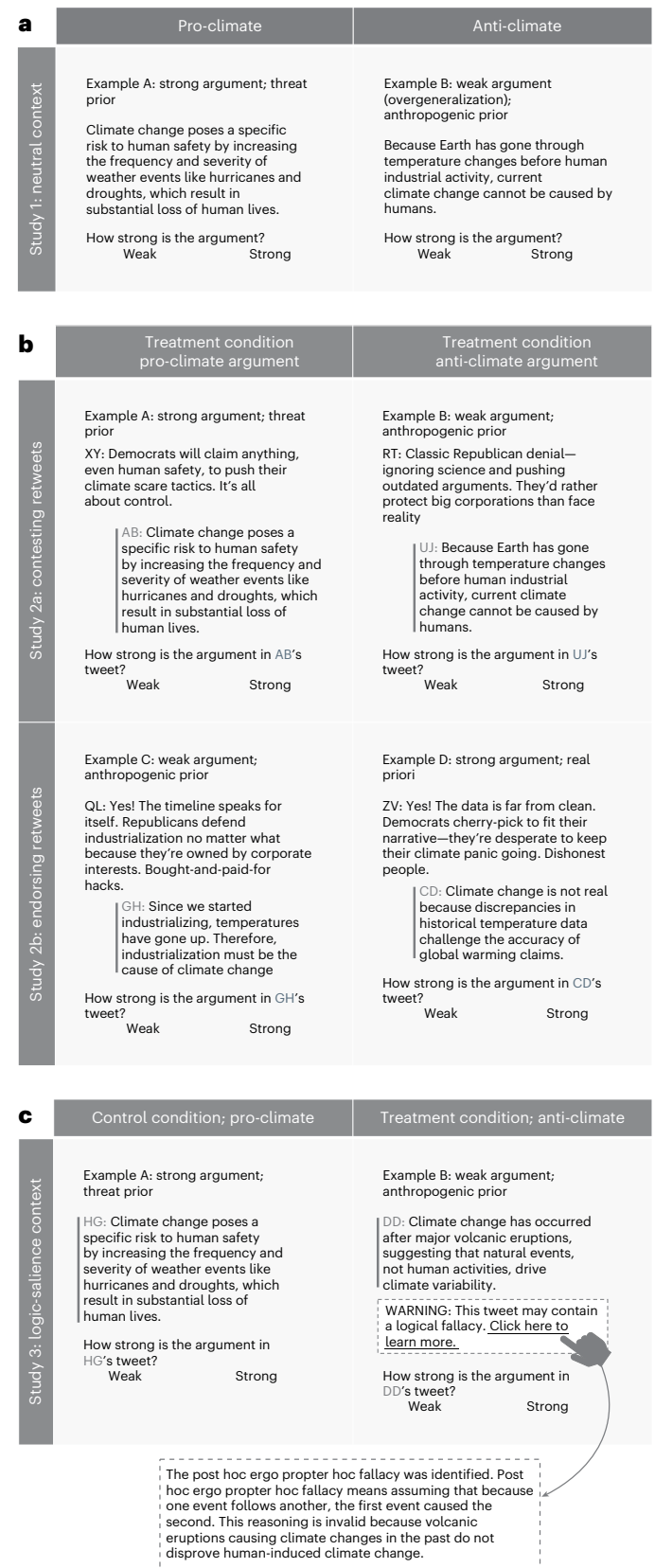


Fig. 1 | Study design in all studies. **a**, Study 1 presented arguments in a neutral context, without additional political or logical cues. On the left is an example of an argument that is pro-climate, strong and about the threat posed by climate change. On the right is an example of an argument that is anti-climate, weak and about the anthropogenic nature of climate change. **b**, In Studies 2a and 2b, participants were placed in either a treatment or control condition. Those in a treatment condition evaluated arguments both with and without a partisan attack, while those in a control condition were not exposed to partisan attacks. On the left are examples of pro-climate arguments treated with an anti-Democrat, contesting retweet (Study 2a) or an anti-Republican, endorsing retweet (Study 2b). On the right are examples of anti-climate arguments with an anti-Republican, contesting retweet attack (Study 2a) or an anti-Democrat, endorsing retweet (Study 2b). **c**, In Study 3, those in the control condition did not see any warnings, regardless of argument strength (see example A). Those in the treatment condition saw all weak arguments with a warning (as in example B) but strong arguments without (as in example A). If presented with a warning, participants had the option to click the link and learn more about the specific fallacy.

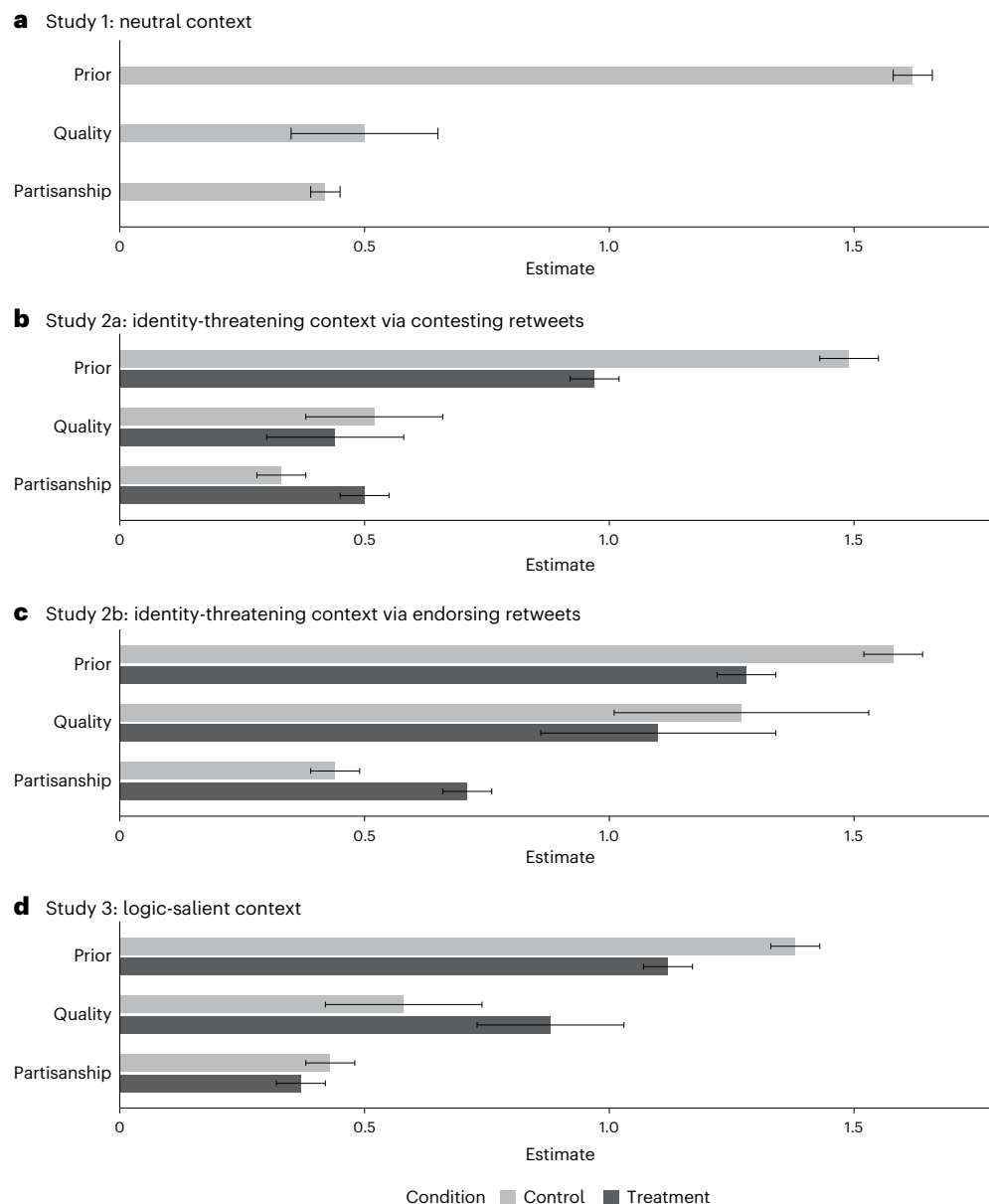


Fig. 2 | Influences of prior beliefs, partisanship and argument quality across contexts. a–d. Bars are standardized beta coefficients that reflect the change in log-odds associated with a one-standard-deviation increase for the primary effects of priors ($pr \times d$), argument quality (a) and partisanship ($pa \times d$) on argument-strength judgements. Positive values indicate higher probability of a ‘strong’ judgement. Error bars represent 95% confidence intervals for the beta estimates. **a**, In a neutral context, participants’ prior beliefs were the strongest predictor. **b, c**, In identity-threatening contexts, the influence of prior beliefs decreased and the influence of political ideology increased for Study 2a (**b**) and

Study 2b (**c**), but not in the expected direction (refer to main text). **d**, In logic-salient contexts, participants were more sensitive to the argument quality. Note, these estimates are standardized fixed-effect estimates from the logistic mixed-effects models. Estimates are model coefficients on the log-odds (logit) scale. To place predictors on a common scale, continuous predictors (prior beliefs, partisanship) were standardized (Z-scored), while dichotomous predictors (argument quality, argument direction) were sum-coded (± 0.5). In Study 1, $N = 1,195$; in Study 2a, $N_{\text{control}} = 469$ and $N_{\text{treatment}} = 452$; in Study 2b, $N_{\text{control}} = 514$ and $N_{\text{treatment}} = 489$; and in Study 3, $N_{\text{control}} = 546$ and $N_{\text{treatment}} = 514$.

treated with caution until tested with a priori studies. Extended analyses reveal broad consistencies in the effect of threat across subgroups (Extended Data Figs. 2 and 3) and fallacy types; however, as for the post hoc analyses, we caution readers not to overinterpret those findings (Supplementary Tables 18–25 and 32–35 and Supplementary Figs. 7–10).

Minimal warnings increase quality discernment

Studies 1, 2a and 2b showed a consistent but weak effect of argument quality. This suggests that people are sensitive to quality, but some may underweight quality cues or lack the mindware to recognize logical fallacies. We, therefore, introduced a logic-salience intervention inspired by social media fact-checking in which minimal warnings are

provided with options to access more information^{40–42}. However, rather than fact-checking, we implement logic-checking—a scalable strategy that places less pressure on platforms to adjudicate politically sensitive, factual or ideological claims. For participants in the treatment condition, a clickable artificial intelligence (AI)-generated warning accompanied weak arguments (Fig. 1c). Clicking gave participants more information about the fallacy. The intervention was deliberately designed to test whether brief warnings—without extensive training or elaborate interventions—could shift argument evaluation.

If making fallacies explicit enhances their impact on evaluation, then the modest influence of argument quality in earlier studies is likely to reflect limited salience (or lack of knowledge) rather than

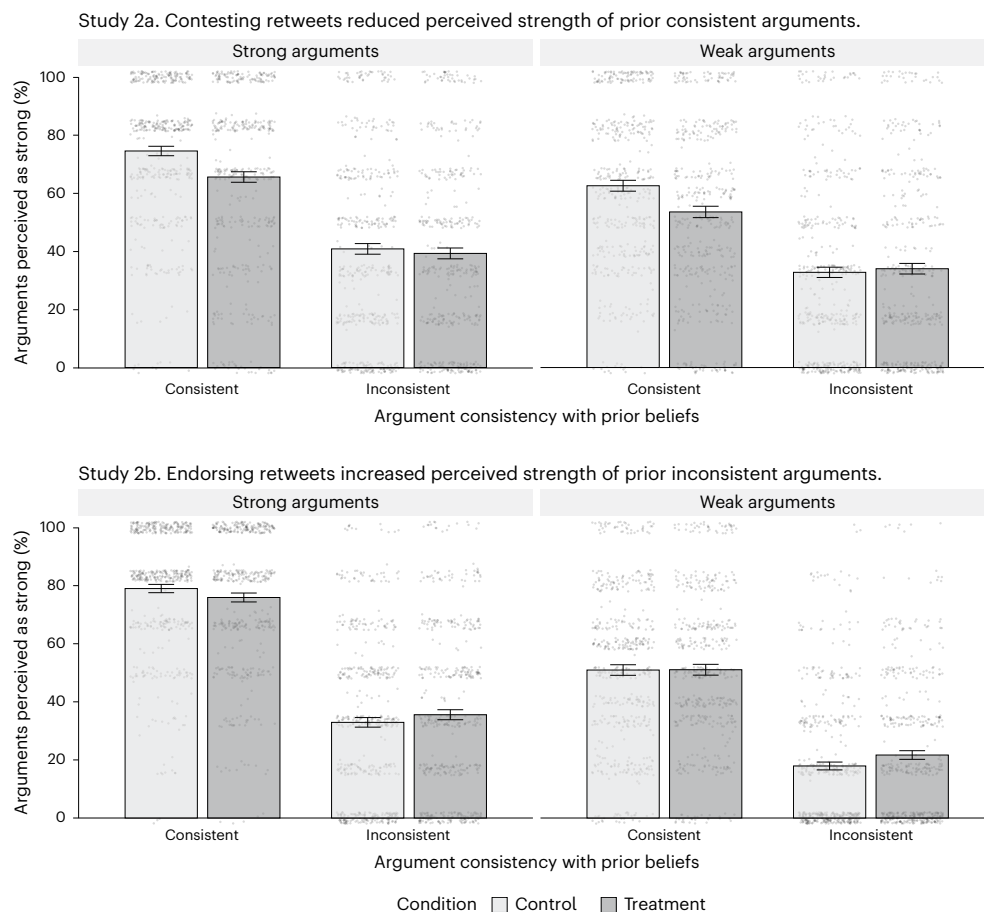


Fig. 3 | Identity threat leads to ‘demotivated’ responding. In Study 2a, consistent arguments in politicized contexts (dark bars) were perceived as strong less often than consistent arguments in neutral contexts (light bars). In Study 2b, inconsistent arguments in politicized contexts were perceived as strong more often than inconsistent arguments in neutral contexts. Dark bars represent

averages for the treatment condition (Study 2a: $N = 452$; Study 2b: $N = 514$), and light bars represent averages for the control condition (Study 2a: $N = 469$; Study 2b: $N = 489$). Error bars represent 95% confidence intervals. Each point is one participant’s mean (the percentage of arguments in that cell that they rated as strong), jittered.

motivated resistance. Further, if warnings enhance quality discernment, regardless of belief consistency, this 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, this would suggest deeper motivated barriers to argument evaluation.

In line with Studies 1, 2a and 2b, prior beliefs remained the strongest predictor of argument evaluation ($\text{pr} \times d: \beta = 2.40, P < 0.001$), while partisan identity ($\text{pa} \times d: \beta = 0.73, P < 0.001$) and argument quality ($a: \beta = 0.50, P = 0.002$) showed smaller but significant effects. The warning treatment had no significant main effect on perceived argument strength, $\beta = -0.13, P = 0.165$, indicating that warnings did not indiscriminately increase scepticism towards all arguments.

We then tested whether these effects varied under the warning treatment. We found no evidence that treatment decreased the effect of partisanship on argument evaluation ($\text{treatment} \times (\text{pa} \times d): \beta = -0.17, P = 0.165$). Partisanship maintained its influence across conditions, suggesting quality cues and partisanship operate independently. In contrast, the intervention decreased reliance on prior beliefs ($\text{treatment} \times (\text{pr} \times d): \beta = -0.44, P < 0.001$) and enhanced argument discernment ($\text{treatment} \times a: \beta = 0.26, P = 0.002$). This was driven by a 4-percentage-point increase in judging weak arguments as weak, and a 1-percentage-point increase in judging strong arguments as strong (Fig. 4). Post hoc analyses revealed that for prior-consistent arguments this was driven largely by reduced acceptance of weak arguments, and

for prior-inconsistent arguments by increased acceptance of strong ones (Supplementary Tables 27–30 and Supplementary Fig. 11).

Exploratory subgroup analyses indicated these gains were broadly consistent: the discernment improvement did not differ significantly by partisan alignment ($\beta = -0.12, P = 0.616$) or belief consistency ($\beta = 0.14, P = 0.565$). The improvement held for Republicans ($\beta = 0.32, P < 0.001$) and Democrats ($\beta = 0.21, P = 0.017$), and for climate believers ($\beta = 0.30, P < 0.001$), with a marginal effect in the same direction for sceptics ($\beta = 0.23, P = 0.054$). The effect also held for politically neutral items ($\text{treatment} \times a: \beta = 0.29, P = 0.009$) and across gender, age and education (Supplementary Table 31). Finally, treatment also reduced perceived strength of weak, but not strong arguments among the neutral arguments ($\text{treatment} \times a: \beta = 0.30, P = 0.002$; Supplementary Table 16 and Supplementary Fig. 12). These findings suggest logic warnings reflect a general mechanism rather than a climate-specific one. Analyses by fallacy type also suggest treatment decreased perceived strength of all fallacy types ($\beta < -0.24, P < 0.01$), but overgeneralization was not significant in the climate item set ($\beta = -0.06, P = 0.576$; Supplementary Tables 32–35). Extended results including full models are provided in Supplementary Tables 27–35.

Understanding what drives these improvements matters both theoretically and for intervention design. First, it is important that participants are not simply downvoting any argument with a warning. The findings suggest that this is not the case: a mechanical downvote would have driven warned arguments uniformly towards the floor, but the warned arguments were still judged as strong a substantial

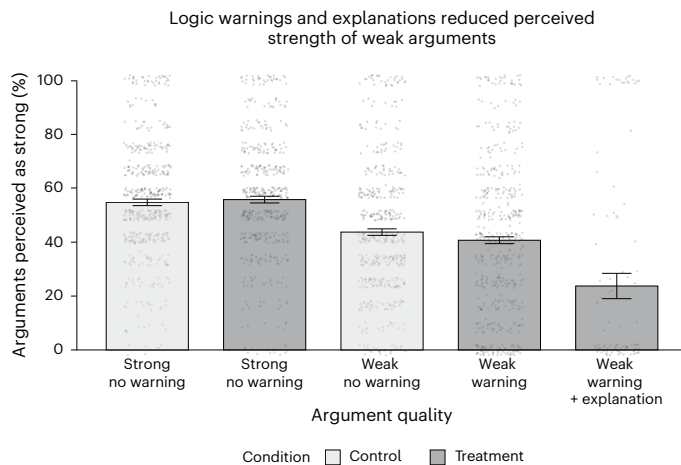


Fig. 4 | Logical warnings increased discernment between weak and strong arguments. Percentage of people perceiving arguments as strong for strong and weak arguments when they were presented with or without warnings. Dark bars represent averages for the treatment condition ($N = 514$), and light bars represent averages for the control condition ($N = 546$). Error bars represent 95% confidence intervals. Each point is one participant's mean (the percentage of arguments in that cell they rated as strong), jittered.

share of the time, and the effect was not uniform but driven mainly by reduced acceptance of belief-consistent weak arguments. Second, warnings alone produced around 5% discernment gains; however, this increased to 21% for participants who accessed the explanations. Although self-selection precludes causal inference, this is consistent with evidence that fact-checking is most effective when accompanied by content-specific explanations⁴³. Interventions that embed explanations may be more effective.

Discussion

Divided climate beliefs impede collective action, and people's ability to discern strong from weak arguments is central to forming accurate ones. Across four studies, prior beliefs emerged as the strongest predictor of argument evaluation, identity-protective accounts found no support and sensitivity to argument quality proved robust across political and belief profiles. These converging observations directly informed a scalable, politically neutral intervention that measurably improves climate argument evaluation.

While we found no evidence for identity-protective reasoning, this does not mean that identity or identity protection could not operate under some other (currently unspecified) form or that partisanship is irrelevant. It is possible, for example, that the aggregate results in Study 2 masked a small identity-protective effect. If the treatment engaged identity protection and social-signal following in opposing directions, the two could partly cancel each other out. However, this would require the identity-protective effect to be so weak as to be able to be counteracted by the social effect, diminishing the identity-protective claim. Nevertheless, the contexts we created may have caused people to 'de-identify' with their political group, potentially giving way to following social cues, and identity protection may still operate under conditions we did not create. Establishing those conditions, and specifying the theory to predict when it should dominate, is a task for future work. Regardless of the role of identity and identity protection, partisanship played a small but consistent role. Partisanship remained associated with argument evaluation even after controlling for prior beliefs, suggesting it may exert a weak independent influence on reasoning. Future work should determine whether this reflects (a) identity protection confined to a minority of participants, undetectable in aggregate results or (b) a distinct cognitive mechanism, separate from identity protection, through which partisanship operates.

Overall, our findings fit a softer coherence-based, Bayesian view. Prior beliefs do not dominate to the extent that argument quality is ignored. Rather, people integrate multiple sources of evidence, including their prior beliefs, argument quality and social signals, which could all fit within flexible Bayesian views²³.

The effect of logic warnings is consistent with the reasoning literature that shows people detect argument quality²¹ and reasoning accuracy can improve with minimal cues⁴⁴. However, how warnings improve argument discernment remains an open question. Warnings flag a problem, name the fallacy and explain why the reasoning fails. This could affect fallacy detection, fallacy understanding or trigger effortful thinking. These routes carry different design implications: if the effect is driven by understanding, warnings should carry explanations; if it is only driven by attention or deliberation, even a contentless flag might suffice. Disentangling them—for instance, by contrasting bare flags with explained ones or by manipulating processing time—is an important next step for maximally effective interventions.

Argument-quality-based interventions offer two practical advantages over fact-checking: they are politically neutral, targeting argumentative structure rather than contested claims, and they are scalable, as AI systems can now detect fallacies without the cost and bias concerns of centralized fact-checking. Warnings improved discrimination equally for Democrats and Republicans with passive exposure yielding around 5% gains and engagement with explanations yielding 21%. This is noteworthy, given that most anti-climate arguments rely on identifiable fallacies³⁵. Embedding explanations directly, via AI or community-notes-style crowdsourcing is, therefore, a promising next step.

Three limitations bound our conclusions. First, we studied only a subset of the cues people encounter online (identity threat and logic salience); source credibility, factual corrections and network signals (likes, shares) remain untested. These may affect reasoning outcomes and bound treatment effectiveness. Logic warnings should be tested in information-heavy environments with many competing cues. Second, climate misinformation extends beyond fallacious arguments to bare assertions, cherry-picked facts and fake-expert appeals containing factual rather than logical errors, which may require complementary methods (fact-checking, source credibility labels or AI systems to flag factual inaccuracies). Relatedly, our findings may not generalize to policy disagreements. Third, different competencies underlying fallacy detection (for example, deductive reasoning, domain knowledge and statistical intuition), it is not clear which competencies our intervention targets.

These findings challenge the prevailing emphasis on identity-protective cognition and single-factor accounts of climate argument evaluation. People retain the capacity to distinguish strong from weak arguments even in polarized contexts, and that capacity can be meaningfully improved through simple cues rather than costly individual persuasion. Climate argument evaluation reflects context-sensitive reasoning in which multiple factors operate in parallel, shifting in relative influence as cues change. The critical question is not which factor dominates, but under what conditions each gains or loses influence, and how those conditions can be designed.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41558-026-02754-2>.

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. *Next 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. & 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**, 121 (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. USA* **115**, 9714–9719 (2018).
12. Kahan, D. M. The politically motivated reasoning paradigm, part 1: what politically motivated reasoning is and how to measure it. In *Emerging Trends in the Social and Behavioral Sciences* (eds Scott, R. A. & Kosslyn, S. M.) 1–16 (Wiley, 2016).
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., Planali, 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. Bavel, J. J. V. & Pereira, A. The partisan brain: an identity-based model of political belief. *Trends Cogn. Sci.* **22**, 213–224 (2018).
18. Taber, C. S. & Lodge, M. Motivated skepticism in the evaluation of political beliefs. *Am. J. Polit. Sci.* **50**, 755–769 (2006).
19. Bago, B., Rand, D. G. & Pennycook, G. Reasoning about climate change. *PNAS Nexus* **2**, pgad100 (2023).
20. Cook, J. & Lewandowsky, S. Rational irrationality: modeling climate change belief polarization using Bayesian networks. *Top. Cogn. Sci.* **8**, 160–179 (2016).
21. 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).
22. Gerber, A. & Green, D. Misperceptions about perceptual bias. *Annu. Rev. Polit. Sci.* **2**, 189–210 (1999).
23. Hahn, U. Argument quality in real world argumentation. *Trends Cogn. Sci.* **24**, 363–374 (2020).
24. Hahn, U. & Harris, A. J. L. What does it mean to be biased: motivated reasoning and rationality. In *Psychology of Learning and Motivation* Vol. 61 (ed. Ross, B. H.) 41–102 (Elsevier, 2014).
25. Mercier, H. The argumentative theory: predictions and empirical evidence. *Trends Cogn. Sci.* **20**, 689–700 (2016).
26. 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).
27. De Neys, W. Advancing theorizing about fast-and-slow thinking. *Behav. Brain Sci.* **46**, e111 (2023).
28. Evans, J. S. B. T. & Stanovich, K. E. Dual-process theories of higher cognition: advancing the debate. *Perspect. Psychol. Sci.* **8**, 223–241 (2013).
29. Johnson-Laird, P. N., Goodwin, G. P. & Khemlani, S. S. Mental models and reasoning. In *International Handbook of Thinking and Reasoning* (eds Ball, L. J. & Thompson, V. A.) 346–365 (Routledge, 2017).
30. Thompson, V. A., Prowse Turner, J. A. & Pennycook, G. Intuition, reason, and metacognition. *Cogn. Psychol.* **63**, 107–140 (2011).
31. Marin, P. & Svedholm-Häkkinen, A. M. Do task instructions affect argument evaluation when accounting for individual differences in reasoning ability and dispositions? *Think. Reason.* <https://doi.org/10.1080/13546783.2026.2670995> (2026).
32. Celniker, J. B. & Ditto, P. H. Of preferences and priors: Motivated reasoning in partisans' evaluations of scientific evidence. *J. Pers. Soc. Psychol.* **127**, 986–1011 (2024).
33. Cook, J., Ellerton, P. & Kinkad, D. Deconstructing climate misinformation to identify reasoning errors. *Environ. Res. Lett.* **13**, 024018 (2018).
34. Lewandowsky, S. Climate change disinformation and how to combat it. *Annu. Rev. Public Health* **42**, 1–21 (2021).
35. Zanartu, F., Cook, J., Wagner, M. & García, J. A technocognitive approach to detecting fallacies in climate misinformation. *Sci. Rep.* **14**, 27647 (2024).
36. Bolsen, T., Druckman, J. N. & Cook, F. L. The influence of partisan motivated reasoning on public opinion. *Polit. Behav.* **36**, 235–262 (2014).
37. Boissin, E., Caparos, S., Voudouri, A. & Neys, W. D. Debiasing system 1: training favours logical over stereotypical intuiting. *Judgm. Decis. Mak.* **17**, 75 (2022).
38. 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).
39. Suhay, E., Bello-Pardo, E. & Maurer, B. The polarizing effects of online partisan criticism: evidence from two experiments. *Int. J. Press Polit.* **23**, 95–115 (2018).
40. 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).
41. 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).
42. 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. USA* **118**, e2104235118 (2021).
43. Drolsbach, C. P., Solovev, K. & Pröllochs, N. Community notes increase trust in fact-checking on social media. *PNAS Nexus* **3**, pgae217 <https://doi.org/10.1093/pnasnexus/pgae217> (2024).
44. Pennycook, G. & Rand, D. G. Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nat. Commun.* **13**, 2333 (2022).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with

the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2026

Methods

Study 1

Participants. In total, 1,631 participants started the experiment and gave consent, and 115 were excluded 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. Thus, altogether we collected data from 1,195 participants (639 female, 545 male and 8 identified as transgender or non-binary; mean age = 48.4 years, s.d. = 17.1 years); 627 participants identified as Republicans and 568 as Democrat. Participants were recruited with Lucid, aiming for national representation for the USA (quota-matched for age, gender and ethnicity). They were paid US\$1.50 for participation. Qualtrics was used to administer the study.

Materials. Reasoning problems. We created 96 reasoning problems, out of which 48 were about climate change and 48 about less-politicized science or technology topics—namely, risks posed by AI, genetically modified organisms and nanotechnology. For climate change arguments, we manipulated two aspects: (1) whether they were pro or contra climate change (directionality; note, pro-climate change arguments affirm the idea that climate change is happening, anthropogenic or risky, and vice versa for contra climate change arguments) and (2) whether they were strong or weak (strength). For 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. For 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 or 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 Zenodo⁴⁵. Altogether, we have four arguments in each category: four strong, arguing that climate change is happening (pro), four strong, arguing that climate change is not happening (contra), and so on. From each of these categories, participants were randomly presented with two problems. Similarly, we applied the same manipulations to the non-politicized arguments, whereas these statements argued for or against the riskiness of AI, riskiness of genetically modified organisms 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 judgement?’, selecting from a scale of 1 (very easy) to 7 (very difficult). We asked for a binary ‘strong’ or ‘weak’ judgement rather than a graded rating, for three reasons. First, it approximates the rapid, categorical judgements people make when evaluating arguments in low-effort settings such as social media; experimental evidence indicates that binary response formats elicit faster and less effortful processing than continuous scales⁴⁶, consistent with our aim of capturing snap judgements rather than deliberate calibration. Second, a binary outcome lets us express treatment effects in directly interpretable percentage-point terms—an effect size that practitioners can readily use. Third, this format follows the conventions of the two literatures closest to our task: misinformation and fake news detection⁴⁰ and reasoning research, where normative argument evaluation is conventionally treated as a categorical judgement^{10,21}.

Prior beliefs. We measured prior beliefs with three different questions according to the argument contents. As arguments were three-fold—(1) climate change is happening, (2) climate change is caused by human activity and (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 or 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 or 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 or 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 posing a threat to agricultural productivity.

Political partisanship. We measured political partisanship with a simple question taken from ref. 19: ‘Which of the following best describes your political preference?’ with the following 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 seven-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. Participants completed the CRT for exploratory purposes only; they are not considered in the current article. We note that while it is possible that completing the CRT may have had some impact on participant responses to the arguments, the consistency of results between the Study 1 and Studies 2a and 2b (without the CRT) with and without the CRT suggest it did not impact the results. Next, participants were instructed that they would see a series of statements about science and technology and would judge how strong each argument was, choosing between ‘strong’ and ‘weak’, followed by a difficulty rating. Statements were presented one at a time in randomized order. The full instructional text is provided in Supplementary Information. Participants were then presented with the selected statements in a randomized order, one at a time. The study was administered with Qualtrics.

Statistical analysis. We used logistic mixed-effects models to analyse the data and applied maximal random-effects 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 centred on 0, on a scale from -1 (climate sceptic) to +1 (climate believer). Partisanship was also centred on 0, and values were transformed to a scale ranging from -1 (strong Republicans) to +1 (strong Democrats). All analyses were conducted with R software version 4.1.1.

Studies 2 and 3

Participants. In Study 2a, 1,832 participants started the experiment and gave consent, while 147 were excluded 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 ($n = 758$); thus, altogether we collected data from 920 participants (419 female, 492 male and 9 identified as transgender or non-binary; mean age = 47.6 years, s.d. = 16.1 years); 421 participants identified as Republicans and 499 as Democrats. Participants were recruited with Lucid and were representative for the 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 US\$1.5 for their participation.

In Study 2b, 1,019 participants started the experiment and gave consent, while 15 were excluded due to giving the incorrect attention-check answer (and, hence, produced no data). Finally, due to overload on the experiment, 1 person dropped out; thus, altogether we collected data from 1,002 participants (493 female, 487 male and 21 identified as transgender or non-binary; mean age = 45.6 years, s.d. = 16.2 years); 386 participants identified as Republicans and 616 as Democrats. Participants were recruited with Prolific and were representative for the US population (quota-matched for age, sex, ethnicity). In total, 514 participants took part in the control group and 489 in the treatment condition (identity-threat manipulation). They were paid £1.99 (approximately US\$2.70) for their participation.

In Study 3, 1,599 participants started the experiment and gave consent, while 189 were excluded due to giving the incorrect attention-check answer (and, hence, produced no data). Altogether, we collected data from 1,060 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, s.d. = 16.7 years); 559 participants identified as Republicans and 501 as Democrats. Participants were recruited with Lucid and were representative for the 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 US\$1.5 for their participation.

Materials and 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 min), we decreased the number of arguments people got from the non-politicized category; they only saw 12 non-politicized items and 24 climate change items. As noted in Fig. 1, the main question was changed from 'How strong is the argument?' in Study 1 to 'How strong is the argument in [AA]'s tweet?' in Studies 2a, 2b and 3. This ensured participants were reporting their assessments of the target arguments and not the surrounding text that was added for the treatment conditions.

Identity-threatening context via contesting retweets (Study 2a). Participants were randomly assigned between subjects to the control (as in 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 from retweet text, above the target argument (see Fig. 1 for examples). To ensure participants 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 towards either Republicans or Democrats but provided no logical counter-argument to the content of the target argument—that is, the retweet did not contain any information that would explicitly help the participant identify a logical flaw in the reasoning of the tweet, if there was one. 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 a political lens, thus prompting them for political identity protection¹⁸. 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 quoted retweets. The full list of argument-retweet matching can be found on Zenodo⁴⁵. 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.'

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

Before the experiment, participants in the partisan cues condition were instructed that they would judge the strength of 36 statements presented as tweets, each accompanied by a re-share comment shown alone for five seconds beforehand; control participants received the Study 1 instructions. The full instructional text is provided in Supplementary Information. The study was administered with Qualtrics.

Identity-threatening context via endorsing retweets (Study 2b). In Study 2a, all pro-climate arguments were paired with Democrat denigration retweets, and anti-climate arguments were paired with Republican denigration retweets. This introduces a confound such that only prior-belief-consistent arguments were paired with identity-attacking retweets. In Study 2b, we reversed this, such that pro-climate arguments were paired with Republican denigrations, and anti-climate arguments were paired with Democrat denigrations, hence creating a set-up where identity-threatening retweets were paired with belief-inconsistent arguments. The full list of argument-retweet matching can be found on Zenodo⁴⁵. These hypothetical retweets always piled on the target arguments and were always relevant to the target argument's content. For example:

Retweet: 'Exactly! Disasters are all over the news. Republicans are too dishonest and too bought by oil money to ever admit it. Bought-and-paid-for cowards.'

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

The experiments were otherwise the same as in Study 2a.

Logical warnings (Study 3). Participants were randomly assigned between subjects to the control (as in Study 1) or the logical cues condition. In the logical cues condition, participants were informed that they may receive AI-generated warnings, and 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 rather than a direct description of the content-based fallacy refutation, to more closely mimic how social media implements such warnings, such as fact-checking information: a warning is placed, but the detailed fact-checking arguments can be reviewed by clicking on the link. Arguments with no argumentative fallacies did not receive any warnings, reflecting the fact that, even though they might not contain argumentative fallacies, they could still be false (just unchecked) or rest on false premises. This avoided participants perceiving these arguments as strong all the time, in a similar way to experiments testing the effectiveness of fact-checking labels on news discernment⁴⁰. These warnings were generated by GPT-4o. In addition to adding ecological validity, this design allowed for additional tests to rule out potential demand effects. Under a demand effect, one could expect no difference in strength judgements between trials when the pop-up was clicked and those when it was not, as strength judgements would already be at the minimum.

Participants in both conditions were told that some of the 36 statements might contain logical fallacies, given a definition of a fallacy and a worked example (appeal to ignorance) and asked to judge the strength of each statement. Participants in the logical cue condition were additionally told that some of the statements had been checked for fallacies by AI, which generated a short warning and an optional clickable explanation wherever a fallacy was found; they were shown an example trial and instructed to judge the statement itself rather than the warning. The full instructional text for both conditions is provided in Supplementary Information. The study was administered with Qualtrics.

Other measures. The same demographic, partisanship and prior measures were applied as in Study 1. Studies 2a and 2b did not contain the CRT to make the experiment shorter to fit the 10-min Lucid criteria.

Statistical analysis. Statistical analysis followed the same scheme as in Study 1. We used logistic mixed-effects models to analyse the data and applied a maximal random-effects 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 arguments and −0.5 for weak arguments, and direction as −0.5 if an argument is pro and 0.5 if an argument is contra. Prior belief was centred on 0 on a scale from −1 (climate sceptic) to +1 (climate believer). Partisanship was also centred on 0 and values were transformed to a scale ranging from −1 (strong Republicans) to +1 (strong Democrats). Treatment was dummy-coded, with 0 representing the control condition and 1 representing the treatment condition (partisan cues in Studies 2a and 2b and logical cues in Study 3). All analyses were conducted with R software version 4.1.1.

Preregistration. Preregistration for Study 1 was available at <https://aspredicted.org/449q-wh8y.pdf>, preregistration for Study 2a was available at <https://aspredicted.org/vw52nr.pdf>, preregistration for Study 2b was available at <https://aspredicted.org/sb9pq5.pdf> and preregistration for Study 3 was available at <https://aspredicted.org/cq9468.pdf>.

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

Ethics statement

Ethical approval for this study was secured by The Ethics Review Board of Tilburg School of Social and Behavioural Science, Tilburg University, under the reference number TSB_RP1699. Informed consent was obtained from all participants before the experiment.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data are available via Zenodo at <https://doi.org/10.5281/zenodo.21964407> (ref. 45).

Code availability

The code to reproduce analysis is available via Zenodo at <https://doi.org/10.5281/zenodo.21964407> (ref. 45).

References

45. Bago, B. & Purcell, Z. A. Dataset and analysis for: Simple logic warnings improve climate argument evaluation across beliefs and partisanship. *Zenodo* <https://doi.org/10.5281/zenodo.21964407> (2026).
46. Rivera-Garrido, N., Ramos-Sosa, M. P., Accerenzi, M. & Brañas-Garza, P. Continuous and binary sets of responses differ in the field. *Sci. Rep.* **12**, 14376 (2022).

Acknowledgements

We thank J.-F. Bonnefon for their helpful discussion during the review of the paper and G. Pennycook for comments on an earlier version of the paper.

Author contributions

The study was conceptualized by B.B. and Z.A.P., who also developed the methodology and conducted the investigation together. B.B. was responsible for curating the data, conducting the formal analysis and developing the experimental software. Z.A.P. secured the funding for this research and created the visualizations. The original draft of the paper was written by B.B. and Z.A.P., and both authors participated in reviewing and editing the final version of the paper.

Funding

Z.A.P. was funded by Agence National de France (ANR-23-AERC-0006).

Competing interests

The authors declare no competing interests.

Additional information

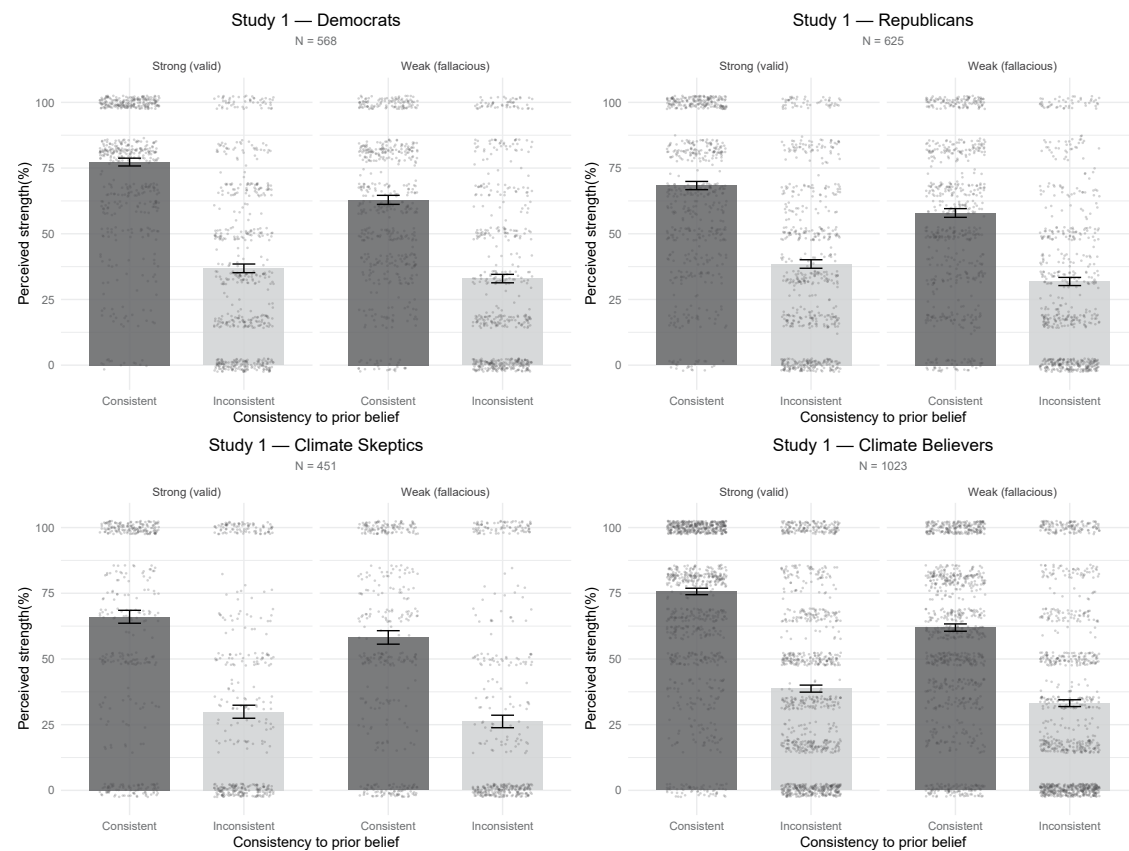
Extended data is available for this paper at <https://doi.org/10.1038/s41558-026-02754-2>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41558-026-02754-2>.

Correspondence and requests for materials should be addressed to Bence Bago.

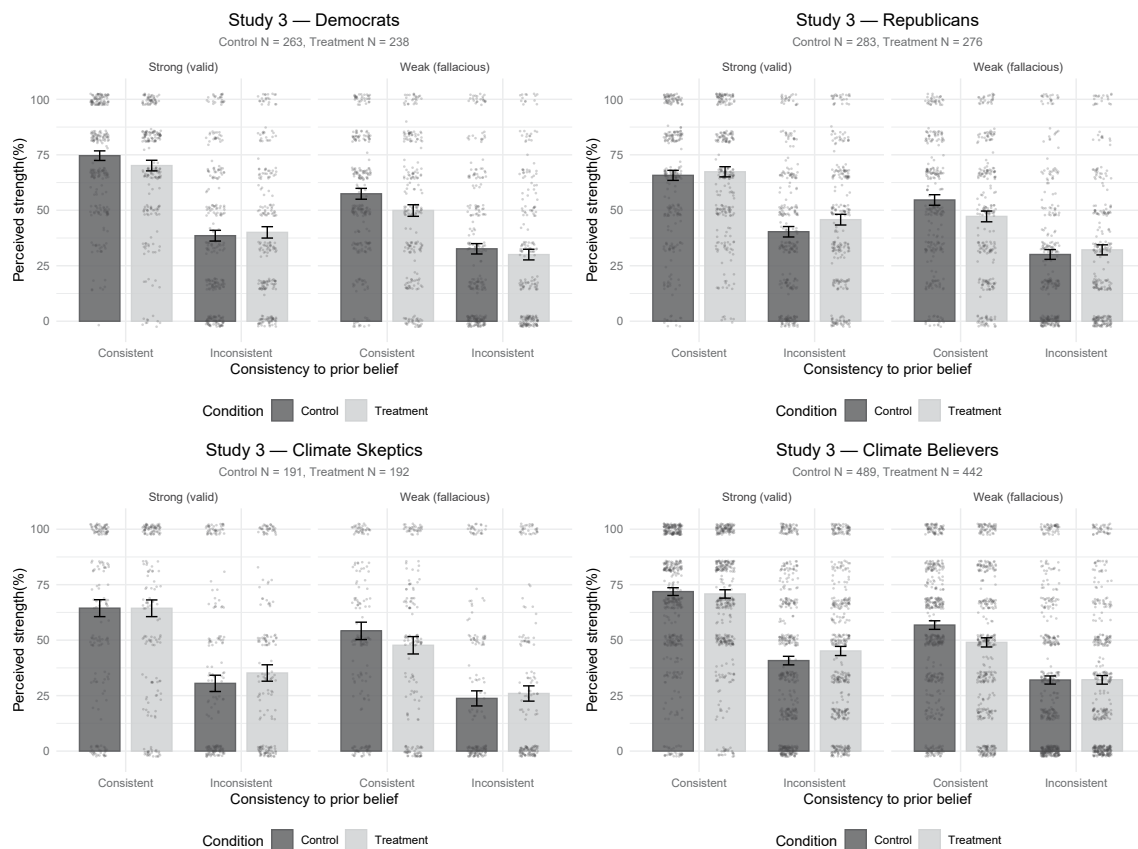
Peer review information *Nature Climate Change* thanks Tobias Brosch, Lea Nahon and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.



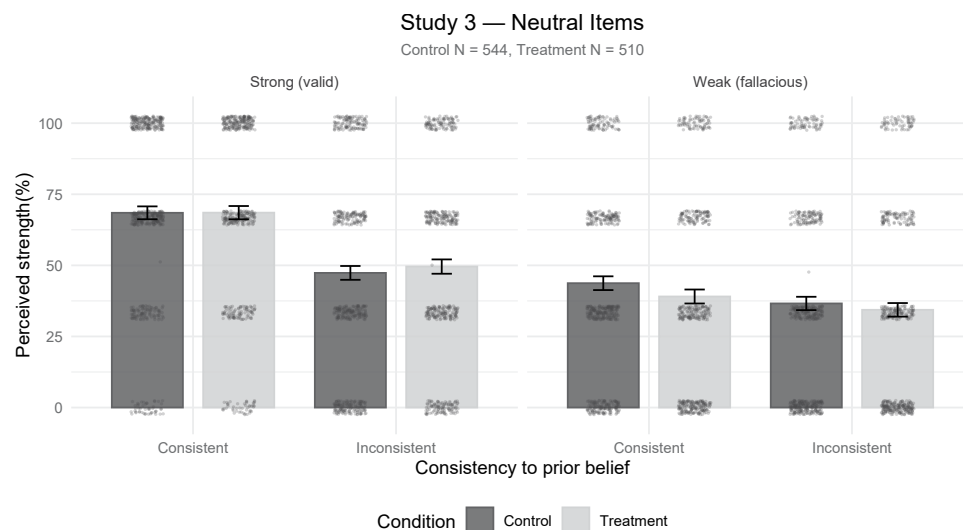
Extended Data Fig. 1 | Strength perception in Study 1 (Climate items), by participant subgroup. Bars show the mean percentage of Climate arguments judged “Strong”. The x-axis contrasts arguments that are consistent versus

inconsistent with the participant’s prior climate belief; panels separate strong (valid) from weak (fallacious) arguments. Study 1 included no experimental manipulation. Error bars are 95% confidence intervals, N = 1195.



Extended Data Fig. 2 | Strength perception in Study 3 (Climate items), by participant subgroup. Bars show the mean percentage of Climate arguments judged “Strong”, by experimental condition (Control vs. Treatment). The x-axis contrasts arguments consistent versus inconsistent with the participant’s prior climate belief; panels separate strong (valid) from weak (fallacious) arguments. Error bars are 95% confidence intervals. The Democrat and Republican panels are defined at the participant level, but the climate-skeptic and climate-believer panels are defined at the *item* level: each argument is classified according to

the participant’s prior belief on the specific climate dimension that argument addresses (whether climate change is happening, whether it is anthropogenic, or whether it poses a risk). The same participant can therefore count as a believer for some items and a skeptic for others – for example, believing that climate change is happening (believer) but not that it is anthropogenic (skeptic) – and so contributes responses to both panels. Their Ns consequently do not sum to the total sample.



Extended Data Fig. 3 | Strength perception in Study 3 (Neutral items).

Bars show the mean percentage of Neutral arguments judged “Strong”, by experimental condition (Control vs. Treatment), across all participants. The x-axis contrasts arguments consistent versus inconsistent with the participant’s prior belief; panels separate strong (valid) from weak (fallacious) arguments. Error bars are 95% confidence intervals. The Democrat and Republican panels are defined at the participant level, but the climate-skeptic and climate-believer panels are defined at the *item* level: each argument is classified according to

the participant’s prior belief on the specific climate dimension that argument addresses (whether climate change is happening, whether it is anthropogenic, or whether it poses a risk). The same participant can therefore count as a believer for some items and a skeptic for others – for example, believing that climate change is happening (believer) but not that it is anthropogenic (skeptic) – and so contributes responses to both panels. Their Ns consequently do not sum to the total sample.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Data analysis

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Data is available on GitHub: https://github.com/bencebago/climate_argumentation

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

We collected a sample that is representative for gender, age and ethnicity from Lucid (Study 1,2a and 3). Gender based analysis was not included, but gender distributions were reported. Gender was measured using self-report questionnaire. For Study 2b, we collected a quota matched sample representative for the US population for sex, age, and ethnicity. Sex was measured as self-report questionnaire.

Reporting on race, ethnicity, or other socially relevant groupings

We collected data on political partisanship, but no other socially relevant grouping. We specifically asked participants: 'Which of the following best describes your political preference?' With 5 response options: 1. Strongly Democratic; 2. Democratic; 3. Lean Democratic; 4. Lean Republican; 5. Republican; 6. Strong Republican.

Population characteristics

See below (Research sample)

Recruitment

Participants for Studies 1,2a,3 were recruited through Lucid, and received \$1.5 for participation. Participants for Study 2b were recruited through Prolific and received £1.99 for participation. Participants could decide freely to engage with the survey given a title and a one-sentence description.

Ethics oversight

Ethical approval for this study was secured by The Ethics Review Board of Tilburg School of Social and Behavioural Science, Tilburg University, under the reference TSB_RP1699. Informed consent was obtained from all participants before the experiment.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

All collected data are quantitative. In all experiments, we gave participants arguments about climate change, both pro and contra. Some arguments included argumentative fallacies in them (weak) and others did not (strong). In all experiments, participants had to judge whether an argument was strong or weak, and we manipulated context: Study 1 - no manipulation, Study 2: politicized context; Study 3: logical context.

Research sample

We collected quota-matched samples from the US, representative for gender (Study 1,2a, 3) or sex (Study 2b), age and ethnicity. In Study 1, we collected data from 1195 participants (639 female, 545 male and 8 identified as transgender or non-binary; Mean age = 48.4 years, SD = 17.1 years); 627 participants identified as Republicans and 568 as Democrat. In Study 2a, we collected data from 920 participants (419 female, 492 male and 9 identified as transgender or non-binary, Mean age = 47.6 years, SD = 16.1 years), 421 participants identified as Republicans and 499 as Democrats. In total, 469 participants took part in the control group and 452 in the treatment condition (identity threat manipulation). In Study 2b, we collected data from 1002 participants (493 female, 487 male and 21 identified as transgender or non-binary, Mean age = 45.6 years, SD = 16.2 years), 386 participants identified as Republicans and 616 as Democrats. In total, 514 participants took part in the control group and 489 in the treatment condition (identity threat manipulation). In Study 3, 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), 559 participants identified as Republicans and 501 as Democrats. In total, 546 participants took part in the control group and 514 in the treatment condition (logical cues manipulation).

Sampling strategy

Quota sampling was used. While determining the sample size we took into account that climate skeptics are less likely to be found in online participant pools, so we decided on sufficiently large samples to each studies. No statistical methods were used to predetermine sample size. Target sample sizes were set in advance and preregistered for each study (preregistration links in Methods). Targets were chosen to be comparable to or larger than those in previous work using similar argument-evaluation designs (ref. 19), and to allow stable estimation of individual-difference effects (prior beliefs, partisanship) in logistic mixed-effects models. Each participant contributed 36–48 binary judgments.

Data collection

We used Qualtrics software to collect data, and participants accessed the survey online through the Lucid Theorem platform for Studies 1,2a and 3) or Prolific (Study 2a) platform. No experimenter was present. Participants were blind to the experimental conditions and hypotheses.

Timing

Data for Study 1 was collected between 06.05.2024-25.06.2024; data for Study 2a was collected between 01.11.2024 - 11.11.2024; data for Study 2b was collected between 19.05.2026-22.05.2026; data for Study 3 was collected between 28.03.2025 - 05.05.2025,

Data exclusions	No data exclusion was applied.
Non-participation	In Study 1, 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. 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. 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. In Study 2a, 1019 participants started the experiment and gave consent, while 15 could not participate due to giving the incorrect attention check answer (and hence produced no data). Finally, due to overload on the experiment 1 person dropped out; hence, altogether, we collected data from 1002 participants.
Randomization	In Study 1, there was no randomisation. In Study 2a and 2b, participants were randomly allocated to experimental condition (political context) or control condition (neutral context). In Study 3, participants were randomly allocated to experimental condition (logical context) or control condition (neutral context).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.
Novel plant genotypes	Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.
Authentication	Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.