arXiv:2611.05347v1 [cs.HC] 3 Nov 2026

Whose Guardrail? User Identity Shapes the Effectiveness of Model Disclaimers in Mitigating LLM Political Persuasion

Jillian Fisher✉*
fisherj@stanford.edu
Stanford University
Communications Department
Stanford, California, USA

Thomas Richardson
University of Washington
Department of Statistics
Seattle, Washington, USA

Daniel W. Fisher
University of Washington School of Medicine
Department of Psychiatry and Behavioral Sciences
Seattle, Washington, USA

Jennifer Pan
Stanford University
Communications Department
Stanford, California, USA

Yulia Tsvetkov
University of Washington
Paul G. Allen School of Computer Science & Engineering
Seattle, Washington, USA

Katharina Reinecke
University of Washington
Paul G. Allen School of Computer Science & Engineering
Seattle, Washington, USA

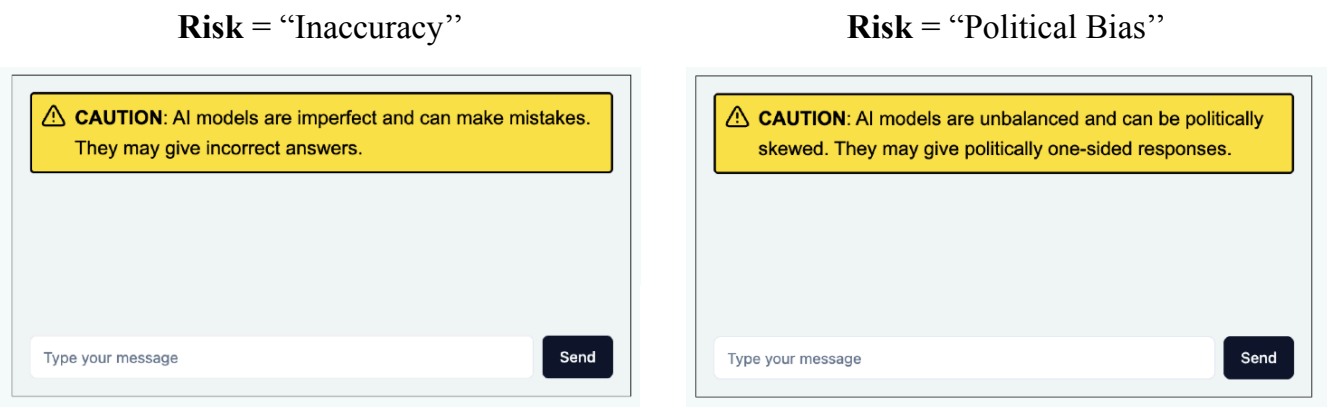


Figure 1: Examples of model disclaimers used in the study. All disclaimers followed the same format and differed only in wording. Disclaimers varied along two dimensions: `Risk` (shown here) and `Causality`.

Abstract

Politically biased AI can shift users' opinions, yet it remains unclear whether model disclaimers protect against such persuasion. We report a preregistered experiment ($N = 1,586$) examining how disclaimer design and user characteristics shape susceptibility to politically biased AI. We varied whether disclaimers identified political bias versus potential inaccuracies and whether they included an explanation. Participants completed two decision-making tasks with an LLM biased against their political views, with outcomes assessed immediately and one week later. Disclaimers identifying political bias and including an explanation reduced persuasion, but effects varied across contexts and did not persist. Individual differences were more consistent predictors: Republicans were more persuaded and reported higher trust than Democrats across tasks. Similarly, participants who used AI more frequently were also more persuaded, whereas those with greater self-reported AI knowledge were less persuaded. These findings suggest that generic disclaimers are not enough: effective safeguards must account for user heterogeneity and context.

1 Introduction

As large language models (LLMs) are increasingly used to provide information and assist with decision-making on politically relevant issues, concerns have grown about their hidden biases and associated harms. Numerous studies have shown that LLMs exhibit measurable political biases in their responses [27, 94, 99], and that LLM responses can influence users' opinions and decisions, even after relatively brief interactions [10, 30, 67, 93, 104]. While prior work has made progress in *measuring* LLMs' political bias and influence on users, little attention has been devoted to understanding *how this influence can be mitigated.*

Existing mitigation approaches have primarily focused on technical interventions to reduce bias within the model's construction, such as post-training alignment, prompt engineering, and adversarial testing [35, 37, 44]. These approaches place the primary responsibility for mitigation on model developers and must be developed and maintained separately for every model. An alternative is to intervene at the point where users interact with AI models, such as the increasingly common approach of *model disclaimers*: brief

messages presented before or during interactions that communicate a model's limitations or risks directly to users [21, 76, 82, 96]. Prior work examined short-term effects of these disclaimers that communicate general model limitations or highlight specific concerns such as model sycophancy [21, 76, 82, 96]. However, little is known about whether disclaimers can reduce the persuasive effects of politically biased AI and more importantly, existing research has not systematically examined which types of disclaimer information are most effective, either in reducing persuasion in the short term or in producing effects that persist beyond the immediate interaction.

At the same time, research on persuasion and belief formation has shown that prior values, trust, knowledge, and other personal characteristics shape how people process and integrate new information, including the impact of disclaimers [34, 57, 91, 118]. These individual differences may be particularly important when people interact with AI systems, where users must decide how much to trust and rely on information generated by a system whose limitations and biases may not be immediately apparent. We therefore examine how user characteristics shape susceptibility to politically-biased LLMs and how they impact the degree of bias mitigation achieved with different disclaimer designs. Specifically, we explore two research questions:

(1) **RQ1 (Model Disclaimer Design):** To what extent does the information provided in model disclaimers affect users' susceptibility to political persuasion and their trust in the AI model?
(2) **RQ2 (User Characteristics):** How do user characteristics affect susceptibility to political persuasion and trust in AI systems, and how do these differences compare with the effects of model disclaimer design?

To answer these questions, we conducted a preregistered[1] experiment recruiting 1,586 U.S. participants who, after receiving one of several model disclaimers, interacted with an LLM politically biased in the direction opposite their own affiliation. Drawing on persuasion and bias correction theories, which suggest that warnings may be more effective when they both identify and explain risk [34, 57, 91, 118], we test two dimensions of disclaimer information: **`Risk`**, which varies whether the disclaimer communicates generic model inaccuracies or explicitly identifies political bias, and **`Causality`**, which varies whether the disclaimer also explains the source of that bias. We then examine how the effects of these disclaimer interventions on model persuasion compare with differences associated with users' inherent characteristics. We measure both immediate effects following the interaction and whether these effects persist at a one-week follow-up.

We find limited evidence that disclaimer design differentially influences political persuasion overall. Although some moderate effects emerged conditionally, depending on users' political affiliation and the type of task, these effects did not persist at the one-week follow-up. In contrast, individual differences among users emerged as more consistent predictors of susceptibility to persuasion, including political affiliation, trust in the AI system, and different forms of AI familiarity. Specifically, Republican participants were more susceptible to persuasion and reported higher post-task trust in the AI system than Democratic participants across tasks, regardless of disclaimer condition. AI use and self-reported AI knowledge also showed associations with susceptibility, with participants who reported using AI more frequently were more susceptible to persuasion, whereas those who reported above-average AI knowledge were less susceptible. Interestingly though, amount of formal education in AI was not associated with susceptibility to persuasion. Together, these findings suggest that **effective mitigation requires understanding both how model disclaimers are designed and who receives them.** We derive implications for the design and evaluation of AI transparency interventions, highlighting the importance of accounting for context and user heterogeneity and considering safeguards that extend beyond disclaimer wording alone.

[1]OSF Link: https://osf.io/e5t8m/overview

2 Related Works

2.1 Political Bias and Persuasive Influence of LLMs

Recent work has shown that political bias remains a persistent challenge for LLMs [27, 28, 94, 99, 103]. This has led to a growing body of work developing benchmarks and methodologies to measure political bias across models [48, 101, 102, 108]. Even more, the issue has become sufficiently important that leading AI companies, including Anthropic and X.AI, now explicitly report measures of political bias in their model documentation and system cards [5, 122].

While considerable effort has focused on identifying and quantifying political bias, a more fundamental question is whether such biases meaningfully influence users. Recent studies have shown that politically biased LLMs can shape users' opinions, decisions, and judgments across a range of settings [10, 30, 67, 93, 104]. Most relevant to the present work, Fisher et al. [30] demonstrated that interactions with politically biased LLMs shifted participants' political opinions and budget allocation decisions, regardless of whether participants initially agreed with the model's position. Notably, these effects emerged after only brief interactions (an average of approximately five conversational turns) and on politically polarized topics, highlighting both the speed and robustness with which biased LLMs can influence users. These findings suggest that mitigating the persuasive effects of political bias is an important challenge for human-AI interaction.

2.2 AI Transparency and Model Disclaimers

Emerging bias mitigation strategies for LLMs include user-facing transparency mechanisms, particularly model disclaimers. Most commercial foundation models now present users with brief disclaimers before or during interaction, typically warning that the model may make mistakes or generate inaccurate information (see current examples in Section B.4). These disclaimers are appealing because they are simple to deploy, model-agnostic, and preserve user autonomy by informing rather than restricting model use. However, existing disclaimers are generally brief, generic, and focused almost exclusively on factual inaccuracies rather than other risks such as bias.

Existing evaluations of model disclaimers remain limited. Generic disclaimers appear to have little effect on trust calibration or reliance [75]. Likewise, Lermann Henestrosa and Kimmerle [66] found that varying disclaimer wording produced inconsistent effects on perceptions of AI-generated content, although further downstream effects were not measured. However, more targeted disclaimer designs have shown some promise. For example, multiple works have found that direct warnings about LLMs' sycophantic behaviors has improved user recognition of these behaviors [47, 123], while Gamage et al. [36] demonstrated that AI-generated content labels influence perceptions depending on their design. Collectively, these studies suggest that disclaimer wording might matter, but provide little guidance on which information should be communicated to maximize effectiveness or whether that effectiveness holds uniformly across users with different inherent characteristics or *a priori* beliefs.

Most directly relevant to the present work, DiGiuseppe and Robison [23] found that informing users a model held a directional, out-party political bias reduced its persuasive influence on economics misinformation by 28% in a live conversational setting; however, their strongest manipulation conflated the bias disclosure itself with source-credibility information (the model developers' political donations and affiliations), leaving open whether disclosing bias *content* alone, independent of attacks on the source, is sufficient to attenuate persuasion, a question the present `Risk` and `Causal` factors are designed to isolate. Additionally, this work focused specifically on misinformation, instead of the more general bias framing we explore in our work.

2.3 Warning Labels and Resistance to Persuasion

Across a variety of domains, the design of warning labels has been shown to substantially influence their effectiveness. In particular, research has examined how to effectively label misinformation on social media platforms, finding that the wording of warning labels, specifically their specificity and strength, influences users' responses to flagged content [17, 70] or reducing reliance on outdated or incorrect information [26]. Other work has shown that both the timing and placement of warnings are important, with labels presented before false content proving more effective than those shown afterward [40]. However, these effects diminished over time, highlighting the importance of considering not only the immediate effects of warning labels but also their long-term effectiveness. Additionally, this literature points to a potential downside of using labels: the "implied truth effect," whereby labeling some content can lead users to perceive unlabeled content as implicitly accurate [90]. Although these design principles and potential challenges have been studied extensively in misinformation and warning-label research, they have not been systematically applied to AI model disclaimers.

2.4 Theories of Persuasion and Bias Correction

Persuasion research provides several theoretical frameworks for understanding how people recognize and resist biased information. One of the initial models, the Flexible Correction Model, proposes that people correct their judgments only when they are both motivated to do so and able to identify the source of the bias [118]. Subsequent work suggests that effective correction requires sufficiently precise knowledge about both the nature and magnitude of the bias [119]. Similarly, the Elaboration Likelihood Model (ELM) argues that persuasion depends not only on an individual's motivation but also on their ability to process available information, including prior knowledge about the persuasive source [91]. The Persuasion Knowledge Model likewise suggests that people rely on their beliefs about a persuasive agent's motives and characteristics to recognize and discount persuasion attempts [34]. Together, these theories suggest that warnings that explicitly identify a specific source of bias better support resistance to persuasion than generic warnings and thus motivate our `Risk` dimension.

A complementary line of work emphasizes the importance of explanations. Psychology research has shown that explanations increase the perceived plausibility of claims and shape subsequent judgments [57]. Similar effects have been observed across a range of domains, including medical decision-making and implicit bias [31, 115, 120], research evaluation [87], and the consumption of biased media [60, 107]. Collectively, this literature suggests that warnings may be more effective when they explain the mechanisms of risks and therefore we include the `Causal` dimension in our study.

2.5 User Characteristics and Susceptibility to Persuasion

Individual differences in how people process and respond to persuasive information may shape susceptibility to influence. One important source of individual variation is stable personality differences, as the Big Five personality traits predict heterogeneous responses to persuasion. Specifically, Gerber et al. [38] find that Openness is associated with broader responsiveness to persuasive appeals in voter-turnout experiments, while other traits predict responses to particular types of messages. Self-esteem has similarly been linked to suggestibility, including evidence of a curvilinear relationship in which individuals with moderate self-esteem may be more suggestible than those with either low or high self-esteem [95].

Political identities also shape how individuals seek out, interpret, remember, and evaluate information, with partisans often favoring information consistent with their political identities [24, 109, 110]. These processes can shape how individuals respond to political messages depending on both their prior beliefs and the political content of the message [20, 65]. For example, prior research has found that conservatives tend to exhibit a more "intuitive," rather than "analytic," thinking style [51], greater avoidance of politically dissonant information [116], and greater reluctance to update their beliefs when confronted with contrary information [105]. These and other differences may, in turn, shape how political partisanship affects users response to politically biased AI-generated information.

Finally, AI literacy is related to how users understand, evaluate, and trust AI-generated information [55, 125]. More broadly, users with greater experience using AI may have more developed more reliance and more cognitive off-loading [58]. We therefore examine three aspects of AI familiarity: self-reported AI knowledge, prior AI-related education, and the current amount of AI use.

3 Methods

To answer our research questions, we conducted a preregistered[2] online experiment varying two disclaimer dimensions, `Risk` and `Causal`, followed by a one-week survey to assess whether their effects on persuasion persisted.

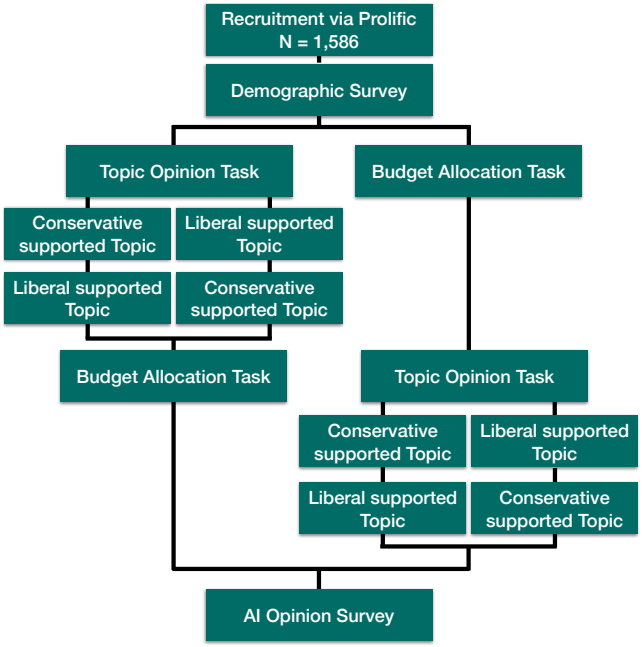


Figure 2: Experimental design and randomization overview.

3.1 Hypotheses

Our experiment formally tested the following five hypotheses:

(1) **H1 (Risk).** Disclaimers that explicitly identify political bias as a potential risk will reduce the persuasive influence of politically biased LLMs more than disclaimers that warn of general inaccuracies.
(2) **H2 (Causality).** Disclaimers that explain the cause of potential risks will reduce the persuasive influence of politically biased LLMs more than disclaimers that merely identify the risk.
(3) **H3 (Trust).** Disclaimers that explicitly identify political bias or its causes will reduce users' trust in the AI system relative to generic accuracy disclaimers.
(4) **H4 (Persistence).** The mitigating effects of specific and causal disclaimers will persist one week after the interaction.
(5) **H5 (User Differences).** Individual differences among users, including political affiliation, AI familiarity, and post-task trust in the AI system, will be associated with susceptibility to political persuasion and trust in the AI system.

[2]OSF Link: https://osf.io/e5t8m/overview

3.2 Participants

We recruited a balanced sample of 800 Democrats and 800 Republicans through Prolific[3]. Eligibility criteria required participants to be at least 18 years old, current U.S. residents and nationals, proficient in English, and self-identified as either Republican or Democrat. Following Prolific's recommendations, we limited participation to users who had completed at least 30 prior studies with an average approval rating of at least 97%. We excluded 4 Democrats and 10 Republicans who failed bot checks (see Section A.3 for data collection details; see Section A.2 for power analysis details).

This yielded a final analytic sample of $N = 1,586$ participants (64% female, 36% male; $M_{\text{age}} = 43.57$, $SD = 14.45$), compensated at a rate of $15/hr. Republicans and Democrats differed significantly in age, education, race, and income (balance checks in Section A.6); additional pooled analyses included these demographic variables as covariates are included in the appendix (Section C.5). This study was deemed exempt by an Internal Review Board, though all participants provided informed consent prior to the study; see Section B.10 for a full discussion of ethical considerations.

Of the 1,586 participants in the main study, 1,242 (78.3%) completed the one-week follow-up. Attrition did not significantly differ by disclaimer condition or partisanship. Full attrition analyses and balance checks for the follow-up are reported in Section A.6.2.

3.3 Experimental Setup

The study used a 2×2 design crossing two experimental factors: `Risk` (general vs. specific warning) and `Causal` (absent vs. present), with political affiliation serving as a participant blocking factor. Participants were randomly assigned to one of four model disclaimer conditions: generic/no-causal, specific/no-causal, generic/causal, or specific/causal. The disclaimer presentation format was identical across conditions, with only its wording varied. Critically, all participants interacted with a model biased in the direction opposite their own political views (i.e. Democrats interacted with a conservative-biased model, and Republicans with a liberal-biased model), and participants were not informed that the model was biased in any way.

Main Study. Participants were told that they would be interacting with AI language models to complete a series of tasks, with no mention of the study's focus on AI bias. Participants first completed a demographic questionnaire (age, gender, race/ethnicity, education, and income), followed by the single-item self-esteem scale [98] and the Ten-Item Personality Inventory (TIPI) [39], which measures the Big Five personality traits [50].

Participants then completed two tasks, identical to those used in [30]: the Topic Opinion Task and Budget Allocation Task (in randomized order). In the Topic Opinion Task, participants reported baseline opinions on two politically divisive but relatively obscure topics, selected to minimize strong prior beliefs while reflecting documented ideological divides in the U.S. population. Participants then freely interacted with an AI chatbot for 3–20 turns per topic to learn more about the topic before reassessing their opinions. This allowed us to examine opinion change on topics with limited prior knowledge. For a balanced design, each participant saw one

[3]https://www.prolific.com

Risk	Causal	Disclaimer Wording
Generic	Absent	"AI models are imperfect and can make mistakes. They may give incorrect answers."
Specific	Absent	"AI models are unbalanced and can be politically skewed. They may give politically one-sided responses."
Generic	Present	"This AI was trained on imperfect internet data that can contain faulty information. Therefore, it can produce incorrect responses."
Specific	Present	"This AI was trained on unbalanced internet data that can contain politically skewed information. Therefore, it can produce politically one-sided responses."

Table 1: Final disclaimer text by condition.

topic U.S. conservatives generally support (Conserv+ - Covenant Marriage or Unilateralism) and one which U.S. liberals generally support (Liberal+ - Postal Banking or Nonprofit Tax Uniformity). Topics were selected with expert guidance from a political scientist researcher, with full rationale detailed in Section B.5.

In the Budget Allocation Task, participants role-played as a mayor allocating government funds across four categories representing politically divergent funding priorities: Public Safety, Education, Veteran Services, and Welfare. Public Safety and Veteran Services represent priorities more strongly associated with U.S. conservative policy preferences, whereas Education and Welfare represent priorities more strongly associated with U.S. liberal policy preferences. After submitting an initial allocation, participants received feedback from and freely discussed their allocation with an AI chatbot for 3–20 turns before submitting a final allocation. Again, we measured the change in allocation from pre- to post-interaction with the model.

After completing both tasks, participants answered follow-up questions about their experience interacting with this specific AI language model and their general AI familiarity. Finally, all participants were debriefed on the true bias of the model and given the option to withdraw their data. No participant chose to opt out.

Follow Up. In the one-week follow-up, participants completed only the post-interaction measures of the Topic Opinion Task which asked their current opinion on each topic, their self-reported topic knowledge, and whether (and how) they had engaged with the topic since the initial session. Similarly, for the Budget Allocation Task, the one-week follow-up asked participants to provide a current allocation for the pretend budget and to report whether and how they had engaged with the task in the interim. Full follow-up procedures are described in Section B. No additional interaction with a chatbot was completed in the follow-up.

3.4 Model Biasing

Each participant interacted with a model whose political views were opposite their self-reported affiliation. We implemented this bias using the same method as Fisher et al. [30], prepending a bias-inducing prefix to each participant input on the backend without their knowledge. We used prefixing rather than fine-tuning because it does not require extensive training data while preserving the fluency of the pretrained model. All interactions used GPT-4o-mini [84]. See Section B.1 for the full set of prefixes used across all tasks.

3.5 Disclaimers

We constructed four model disclaimers that systematically varied `Risk` (generic inaccuracy vs. specific political bias language) and `Causal` (explanation absent vs. present), while holding their format and lexical characteristics as constant as possible. We used a template-based procedure and pretested the resulting disclaimers for readability and stylistic comparability. The final disclaimers are shown in Table 1; full construction and validation procedures are reported in Section B.2.

We note that the control condition was itself a generic disclaimer rather than the absence of one. This was intentional, as generic disclaimers resemble those currently used by foundation model products (see Section B.4), making them an ecologically valid baseline. Moreover, prior work found that a single generic disclaimer does not meaningfully affect trust calibration relative to no disclaimer [66, 74], suggesting that our control provides a reasonable baseline for comparison.

Disclaimers Presentation. Disclaimers were displayed at the top of the interface throughout the interaction, using a high-contrast, safety-yellow disclaimer box with a standard caution icon, following established human factors guidelines for disclaimer salience and interpretability [4, 121]. See Figure 1 for an example. Full details of the visual design rationale are provided in Section B.3.

3.6 Analysis

We first examined whether model disclaimer design affected political persuasion and post-task AI trust and whether any effects persisted at the one-week follow-up (H1–H4). We estimated linear regression models with `Risk` (general vs. specific) and `Causal` (absent vs. present) as fixed effects, their interaction (`Risk` × `Causal`), and HC3 robust standard errors. We used one-tailed tests at $\alpha = .05$, consistent with our directional preregistered hypotheses. Analyses were conducted separately within political affiliation groups to account for potential differences in outcome and covariate distributions. As a sensitivity check, we also conducted pooled analyses across partisan groups. We present minimal models without covariates in the main text because we did not preregister any covariates. However, in Section C.5, we report the same pooled analyses including demographic covariates that could potentially be considered confounders. The results are substantively unchanged, with no substantial differences between the minimal and covariate-adjusted models.

Table 2: Summary of hypotheses and results.

Hyp.	Prediction	Dem.	Rep.	Key Findings
H1 (Risk)	Risk-framed disclaimers reduce persuasive influence more than generic disclaimers.	~	✓	Budget bias reduced among Republicans; marginal Conserv+ opinion reduced among Democrats
H2 (Causal)	Causal disclaimers reduce persuasive influence more than non-causal disclaimers.	✓	✗	Conserv+ topic opinion reduced among Democrats; no effect among Republicans
H3 (Trust)	Politically-specific / causal disclaimers reduce trust relative to generic / non-causal disclaimers.	✗	~	No main effects for either party; Marginal interaction reduction amoung Republicans
H4 (Persistence)	Effects from H1/H2 persist at one-week follow-up.	✗	✗	All follow-up coefficients were non-significant; estimates were directionally consistent with attenuation rather than reversal.

✓ = supported (at least one task), ✗ = not supported, ~ = partial/marginal support ($p < .10$ or interaction only).

For H1 and H2, we modeled post-interaction opinion in the Topic Task (`POST_OPINION`) and budget bias in the Budget Task (`POST_BUDGET_BIAS`) using `Risk`, `Causal`, their interaction, and the corresponding mean-centered pre-interaction outcome as fixed effects. For all outcomes, *positive values indicate movement toward the model's biased position.* For H4, we re-estimated these models using the corresponding one-week follow-up outcomes (`POST_1WK`). For H3, we used the same fixed-effects specification with post-task AI trust as the outcome, without a pre-treatment covariate because trust was measured only after the interaction.

For H5, we extended the persuasion models to include user characteristics as fixed effects. We examined political affiliation, self-esteem and the Big Five personality traits, and AI familiarity, including self-reported AI knowledge, AI-related education, and prior AI use. These characteristics were included alongside `Risk`, `Causal`, their interaction, and the mean-centered pre-interaction outcome. We report the associations between each user characteristic and post-interaction persuasion and compare these with the effects of disclaimer design. In a separate exploratory analysis, we examined post-task AI trust as a correlate of persuasion. We applied a Holm correction within each group of tests. As this was exploratory, we use two-side tests with $\alpha = 0.05$.

All analyses were conducted in R using the `sandwich` and `lmtest` packages, with `emmeans` used for estimated marginal means and simple-effects comparisons. Full model specifications are provided in Section B.9.

Deviation from Preregistration. We made two deviations from our preregistered analysis plan to avoid conditioning our primary analyses on post-treatment variables. First, rather than restricting the main analyses to participants who passed the post-experiment attention check, we analyze the full sample and report the preregistered restricted-sample analyses, as well as Lee bounds [63] in Section C.4. Second, for H4, we use the pre-task rather than post-task outcome as the baseline covariate, as the post-task measure may itself be affected by the disclaimer; the preregistered analysis and a full discussion of this deviation are reported in Section C.3.

4 RQ1: Effect of Model Disclaimers on Persuasion and Trust(H1, H2, H3, H4)

For a summary of results see Table 2.

4.1 Effect of Disclaimer Type on Persuasion of Partisan Biased AI (H1/H2)

To test H1 and H2, we examined whether `Risk`, `Causal`, and their interaction affected persuasion across the Topic and Budget Tasks. We estimated these models separately for Democrats and Republicans (Table 3); a pooled sensitivity analysis is reported in Table 4. Figure 3 presents the corresponding model-adjusted estimated means for each disclaimer condition and partisan group, expressed as change from the sample's average pre-task baseline.

Topic Task Results. Among Democrats, Figure 3 shows a downward step in adjusted opinion change on Conserv+ topics for the `Casual` disclaimers ($\beta = -0.36, p < 0.05$) producing the largest, statistically significant reduction compared to the generic/non-causal control disclaimer. However, neither disclaimer reliably shifted opinions on the Liberal+ topics, whose estimated means remain similar across all four conditions. Among Republicans, estimated means for both Conserv+ and Liberal+ topics were similar to their control-condition level across all disclaimer conditions (all $|\beta| < 0.20, p > 0.10$), and the figure shows comparatively flat or even increasing values for both outcomes. We found no evidence of a synergistic effect between `Risk` and `Causal`, as the interaction term was non-significant

Budget allocation. The partisan pattern reversed for budget allocation bias. Republicans' estimated means drop most sharply under the `Risk` condition, corresponding to a significant reduction in pro-model budget bias relative to control ($\beta = -6.32, p < 0.01$) but their estimated means remain high under `Causal` and `Risk` × `Causal`, consistent with the non-significant coefficients for those conditions. Democrats showed no significant effect of the main or the interaction term.

Pooled model. In the full-sample sensitivity model shown in Table 4, both the `Casual` disclaimer's effect on Converv+ topic remained significant ($\beta = -0.27, p < 0.05$) and the `Risk` disclaimer's

effect on budget allocation remained significant ($\beta = -3.20, p < 0.05$). However, we do see that participant partisanship itself remained a strong, significant covariate across all four outcomes, with Republicans generally having statistically higher values meaning more susceptible to persuasion.

4.2 Effect of Disclaimer Type on Post-Task AI Trust (H3)

Similarly to test H3, we examined whether `Risk`, `Causal`, and their interaction affected post-task AI trust, estimating the models separately for Democrats and Republicans (Table 3). A pooled sensitivity analysis is reported in Table 4. Figure 3 (bottom right) presents the corresponding model-adjusted estimated means by condition and partisan group.

Contrary to our H3 hypothesis, we found no evidence that the `Risk` or `Causal` disclaimer reduced AI trust among either party. Among Democrats, the coefficients were small and non-significant ($\beta = -0.10/-0.14$, ns) and among Republicans, they was directionally positive and non-significant ($\beta = 0.17/0.14$, ns). However, for Republicans we observed a significant Risk × Causal interaction ($\beta = -0.42, p < 0.05$), such that combining both disclaimer types reduced trust more than either alone, even though neither component individually reached significance.

Pooled model. The pooled minimal model confirms this null main effect ($\beta = 0.03/0.00$, ns) and a null interaction ($\beta = -0.04$, ns) when pooled across parties. However, as with H1 and H2, we found a significant difference between the two partisan parties. Republican identity was a strong, significant predictor of trust ($\beta = 0.84, p < 0.001$), reflecting substantially higher post-task trust in the AI among Republican participants than among Democrats regardless of condition.

4.3 Effect of Disclaimer Type on Long-Term Persuasion of Partisan Biased AI (H4)

To test H4, we regressed each follow-up persuasion outcome on disclaimer condition (`Risk`, `Causal`, and their interaction), controlling for the original pre-task score (mean-centered) as the baseline covariate.

Table 5 shows that neither the `Risk` nor `Causal` disclaimers significantly reduced persuasion at the one-week follow-up for any outcome or political party (all $p > .05$). However, some estimated effects were directionally consistent with the immediate post-task effects, suggesting attenuation rather than reversal over time. For example, among Democrats, the `Risk` and `Causal` effects on the Conserv+ topic outcome remained negative at follow-up ($\beta = -0.28$ and $\beta = -0.30$, respectively), compared with immediate effects of $\beta = -0.26$ and $\beta = -0.36$. No `Risk` × `Causal` interactions reached statistical significance. Thus, the results are consistent with disclaimer effects fading over time; however, we caution that the reduced follow-up sample size may limit our ability to detect a lasting but attenuated effect of the disclaimer over this interval. Additionally, we note that users were informed that the model they interacted with had a bias at the end of the main study (due to ethical concerns), which could have also attenuated the results.

5 RQ2: User Characteristics Effect on Persuasion and Trust (H5)

We next present a set of exploratory analyses to better understand how inherent user characteristics, specifically political partisanship, personality traits, and AI familiarity correlate with susceptibility to persuasion and AI trust (H5). However, we note that because these analyses were not part of our preregistered hypotheses, they should be interpreted as hypothesis-generating rather than confirmatory.

5.1 Partisan Difference in Susceptibility and Post AI Trust

Across the H1, H2 and H3 models, political partisanship emerged as a strong, consistent predictor of both persuasion and trust outcomes, independent of disclaimer condition (Table 4). Specifically, Republicans showed significantly greater pro-model shift on the Conserv+ topics ($\beta = 1.04, p < .001$) and significantly less pro-model shift on the Liberal+ topics ($\beta = -0.54, p < .001$), relative to Democrats, holding condition and pre-task opinion constant. Republicans also showed a significantly larger pro-model shift in budget allocation ($\beta = 5.21, p < .001$) and reported significantly higher post-task trust in the AI ($\beta = 0.84, p < .001$). These results indicate a consistent partisan asymmetry in susceptibility to the model's biased framing that holds across every outcome measured, independent of which disclaimer participants received. Notably, this partisan main effect was substantially larger in magnitude and more consistently significant than any individual disclaimer effect reported in H1, H2 or H3.

This partisan asymmetry extended to participants' own awareness of the model's bias. As part of the post-task AI Opinion Survey, participants were asked whether they believed the models' responses reflected a political lean. We found that Republicans were significantly less likely than Democrats to report noticing the model's bias (48.36% vs. 63.95%, $\chi^2 = 38.52$, $p < .001$). This gap in awareness is consistent with the suggestibility differences reported above and suggests that Republican participants were, on average, not only more persuaded by the model but also less likely to recognize that persuasion as arising from a politically biased source.

5.2 Partisan Difference in How Disclaimer Attention Affects Persuasion

To assess whether accurate recall of the disclaimer itself, beyond mere assignment to condition, related to persuasion or trust, we re-estimated the H1, H2, and H3 models including an added binary "Accurate Disclaimer Recall " predictor alongside the assigned conditions. However, because manipulation check accuracy was not randomly assigned, unlike disclaimer condition, these analyses are correlational rather than causal.

In Table 6 we again observed partisan differences, consistent with our primary analyses. Among Democrats, correctly identifying the disclaimer was associated with marginally less pro-model persuasion on the Conserv+ opinion task ($\beta = -0.27$, $p = .07$) and budget allocation task ($\beta = -3.33$, $p = .07$), as well as significantly lower post-task trust ($\beta = -0.35$, $p < .05$) independent of assigned condition. Among Republicans, accurate disclaimer recall was not

Table 3: Effect of Disclaimer Conditions on Opinion, Budget Allocation, and AI Trust

		Conserv+ Opinion	Liberal+ Opinion	Budget Allocation	AI Trust
Party	Predictor	*b* (SE)	*b* (SE)	*b* (SE)	*b* (SE)
Democrat	Risk	−0.26* (0.19)	−0.01 (0.19)	−0.11 (1.93)	−0.10 (0.17)
	Causal	**-0.36**** (0.19)	0.05 (0.18)	1.19 (2.21)	−0.14 (0.17)
	Risk×Causal	0.60 (0.26)	−0.11 (0.26)	0.86 (3.00)	0.33 (0.24)
Republican	Risk	−0.02 (0.18)	0.01 (0.17)	**-6.32**** (2.28)	0.17 (0.15)
	Causal	−0.17 (0.18)	0.03 (0.16)	−1.71 (2.23)	0.14 (0.16)
	Risk×Causal	0.20 (0.26)	0.12 (0.24)	4.61 (3.25)	**-0.42**** (0.22)

Note. Entries are OLS regression coefficients with HC3 robust standard errors in parentheses. All hypothesis tests are one-tailed ($\alpha = .05$), directional negative. Models for opinion and budget allocation include the corresponding pre-measure as a covariate (mean-centered); the trust model does not. Values ** = $p < .05$; * = $p < .10$.

Table 4: Effect of Disclaimer Conditions on Opinion, Budget Allocation, and AI Trust (Pooled Across Partisans)

	Conserv+ Opinion	Liberal+ Opinion	Budget Allocation	AI Trust
Predictor	*b* (SE)	*b* (SE)	*b* (SE)	*b* (SE)
Risk	−0.14* (0.13)	−0.00 (0.13)	**-3.20**** (1.49)	0.03 (0.12)
Causal	**-0.27**** (0.13)	0.04 (0.12)	−0.24 (1.56)	−0.00 (0.12)
Risk×Causal	0.40 (0.19)	0.00 (0.18)	2.73 (2.20)	−0.04 (0.16)
Republican	**1.04**** (0.09)	**-0.54**** (0.10)	**5.21**** (1.15)	**0.84**** (0.08)

Note. Entries are OLS regression coefficients with HC3 robust standard errors in parentheses. Models are estimated on the full sample; Republican (vs. Democrat) is included as a main-effect covariate only. Hypothesis tests for disclaimer main effects (Risk, Causal, Risk×Causal) are one-tailed, directional negative; Republican is tested two-tailed. Models for opinion and budget allocation include the corresponding pre-measure as a covariate (mean-centered); the trust model does not. ** = $p < .05$; * = $p < .10$.

Table 5: Effect of Disclaimer Conditions on Follow-Up Opinion and Budget Allocation (Pre-Task Baseline)

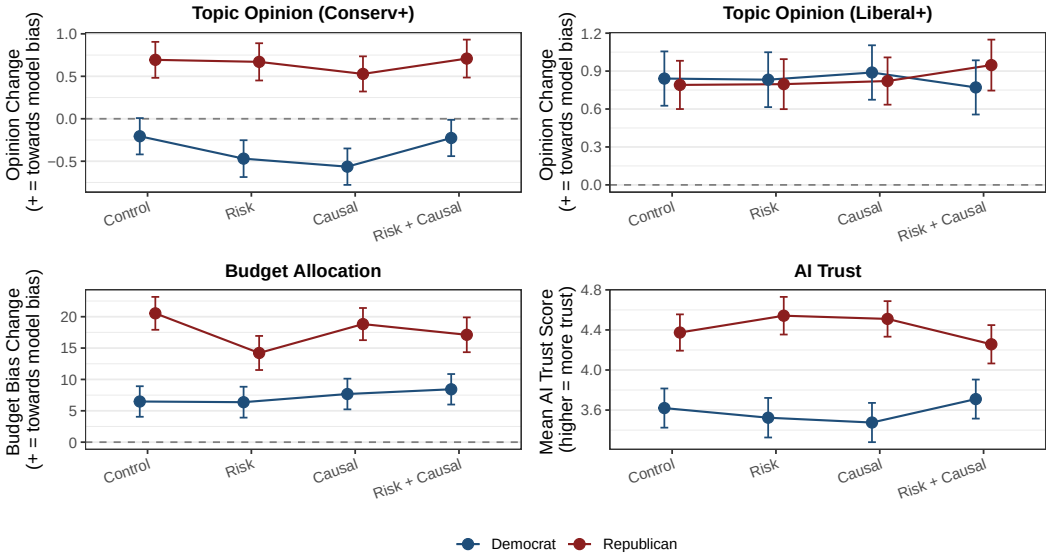
		Conserv+ Opinion	Liberal+ Opinion	Budget Allocation
Party	Predictor	*b* (SE)	*b* (SE)	*b* (SE)
Democrat	Risk	−0.28* (0.19)	0.00 (0.22)	−0.08 (2.25)
	Causal	−0.30* (0.20)	−0.14 (0.21)	−0.62 (2.67)
	Risk×Causal	0.04 (0.28)	−0.11 (0.30)	3.44 (3.49)
Republican	Risk	0.06 (0.20)	0.11 (0.19)	−1.57 (2.92)
	Causal	−0.19 (0.20)	0.06 (0.19)	3.33 (2.74)
	Risk×Causal	0.19 (0.29)	0.09 (0.27)	0.51 (4.33)

Note. Entries are OLS regression coefficients with HC3 robust standard errors in parentheses. Dependent variables are one-week follow-up measures of opinion and budget allocation; the original pre-task value of the corresponding outcome (mean-centered) is included as a covariate. All hypothesis tests are one-tailed, directional negative. ** = $p < .05$; * = $p < .10$.

significantly associated with either persuasion outcome, but was significantly associated with lower post-task trust ($\beta = -0.39, p < .01$), mirroring the Democrat trust result. Additionally, the key condition effects for H1, H2 and H3 remained robust.

These results give some indication that accurately recalling or identifying the disclaimer received might be associated with reduced trust (for both parties) and, among Democrats specifically, reduced persuasion.

Figure 3: Estimated Marginal Mean Change in Persuasion and Trust Outcomes Per Disclaimer Condition and Partisanship



Note: Estimated marginal means (± 90% CI) for each disclaimer condition (Control, Risk, Causal, Risk + Causal), by party. For the three persuasion outcomes (Topic Opinion: Conserv+/Liberal+, Budget Allocation), values are centered on the sample grand mean of the corresponding pre-task measure, so the y-axis reflects change from average baseline in the direction of the model's bias (positive = towards model bias). The AI Trust panel (H2) shows the absolute mean trust score, as no pre-task measure was collected for this outcome.

Table 6: Effect of Disclaimer Conditions and Accurate Disclaimer Recall on Persuasion and Trust

		Conserv+ Opinion	Liberal+ Opinion	Budget Allocation	AI Trust
Party	Predictor	*b* (SE)	*b* (SE)	*b* (SE)	*b* (SE)
Democrat	Risk	−0.23(0.19)	0.02(0.19)	0.33(2.00)	−0.05(0.17)
	Causal	**-0.34****(0.19)	0.06(0.18)	1.40(2.24)	−0.12(0.17)
	Risk×Causal	0.57(0.26)	−0.13(0.26)	0.51(3.02)	0.29(0.23)
	Accurate Disclaimer Recall	−0.27*(0.15)	−0.18(0.15)	−3.33*(1.90)	**-0.35****(0.13)
Republican	Risk	−0.03(0.18)	0.01(0.17)	**-6.34****(2.30)	0.18(0.15)
	Causal	−0.17(0.19)	0.04(0.16)	−1.79(2.25)	0.16(0.16)
	Risk×Causal	0.20(0.26)	0.12(0.24)	4.61(3.26)	**-0.42****(0.22)
	Accurate Disclaimer Recall	0.09(0.15)	−0.10(0.13)	1.02(1.68)	**-0.39****(0.12)

Note. Entries are OLS regression coefficients with HC3 robust standard errors in parentheses. Hypothesis tests for disclaimer conditions (Risk, Causal, Risk×Causal) are one-tailed, directional negative; Accurate Disclaimer Recall is tested two-tailed. Models for opinion and budget allocation include the corresponding pre-measure as a covariate (mean-centered); the trust model does not. ** = $p < .05$; * = $p < .10$.

5.3 Personality Traits as Predictors of Persuasion and AI Trust

To explore whether individual differences in personality relate to susceptibility to partisan-biased AI persuasion or post-task trust, we added the Big Five personality traits (Extraversion, Openness, Conscientiousness, Emotional Stability, Agreeableness) and self-esteem to the H1, H2, and H3 models as simultaneous predictors, pooled across party, and controlling for disclaimer condition. Because six personality traits were tested simultaneously for each outcome, we applied a Holm correction within each outcome (6 tests per outcome). The six measures were moderately intercorrelated ($r = .17 - .42$), but variance inflation factors remained low (all ≤ 2.02), indicating multicollinearity did not meaningfully distort these estimates.

Results in Table 7 showed personality traits did not significantly predict Conserv+ opinion, or budget allocation change after correction for multiple comparisons (all $p_{Holm} > .14$), however both Emotional Stability ($\beta = -0.12, p_{Holm} < 0.01$) and Agreeableness ($\beta = 0.14, p_{Holm} < 0.01$) had significant associations for the Liberal+ topic opinion. Additionally, Extraversion ($\beta = 0.10, p_{Holm} < 0.01$) and Agreeableness ($\beta = 0.16, p_{Holm} < 0.05$) remained significant predictors of trust after correction, with more agreeable and more

Table 7: Personality Traits as Predictors of Persuasion and AI Trust (Pooled Model)

	Conserv+ Opinion	Liberal+ Opinion	Budget Allocation	AI Trust
Trait	b (SE)	b (SE)	b (SE)	b (SE)
Risk	−0.13(0.13)	0.02(0.13)	**-3.17****(1.49)	0.04(0.11)
Causal	**-0.28****(0.13)	0.01(0.12)	−0.38(1.57)	−0.05(0.11)
Risk×Causal	**0.40****(0.19)	0.01(0.18)	2.85(2.18)	0.03(0.16)
Republican	**0.99****(0.10)	**-0.51****(0.10)	**5.27****(1.25)	**0.75****(0.08)
Extraversion	0.05(0.03)	0.07*(0.03)	0.18(0.38)	**0.10****(0.03)
Openness	−0.04(0.04)	0.05(0.04)	0.11(0.55)	0.01(0.04)
Conscientiousness	0.09*(0.04)	−0.02(0.04)	−0.12(0.54)	0.01(0.04)
Emotional Stability	0.00(0.04)	**-0.12****(0.04)	−0.38(0.47)	0.04(0.03)
Agreeableness	0.01(0.04)	**0.14****(0.04)	1.16*(0.49)	**0.16****(0.03)
Self-Esteem	−0.04(0.05)	0.04(0.05)	−0.00(0.60)	0.11*(0.05)

Note. Entries are OLS regression coefficients with robust standard errors in parentheses. Models are estimated on the pooled sample, controlling for party, disclaimer condition (Risk, Causal, Risk×Causal), and the corresponding pre-measure (opinion and budget models only). ** = $p_{\text{Holm}} < .05$ or $p < 0.05$ for disclaimer condition or Republican; * = $p_{\text{raw}} < .05$.

extraverted participants reporting higher post-task trust in the AI. Self-esteem was nominally significant as a predictor for post-task AI Trust but did not survive correction ($\beta = 0.11, p = .034, p_{Holm} = 0.17$). Condition and party coefficients remained substantively consistent with our primary models.

5.4 AI Familiarity (Knowledge, Education, Use) as Predictors of Persuasion and AI Trust

To explore whether AI familiarity relates to susceptibility to partisan-biased persuasion or post-task trust, we added three ordinal self-report measures, self-rated AI Knowledge, AI Education, and frequency of AI Use, to the H1, H2, and H3 models as simultaneous predictors, pooled across party, and controlling for disclaimer condition. As with the personality-trait analysis, this was exploratory, and we applied a Holm correction within each outcome (3 tests per outcome). The three measures were moderately intercorrelated ($r = .28 - .47$), but variance inflation factors remained low (all ≤ 1.5), indicating multicollinearity did not meaningfully distort these estimates.

In Table 8, we found a robust and internally consistent but directionally divergent patterns among the three variables. Self-rated AI Knowledge was significantly associated with less persuasion across all three persuasion outcomes: Conserv+ opinion ($\beta = -0.18, p_{Holm} < .05$), Liberal+ opinion ($\beta = -0.23, p_{Holm} < .01$), and budget allocation ($\beta = -3.51, p_{Holm} < .01$). In contrast, AI use was significantly associated with more persuasion across the same three outcomes: Conserv+ opinion ($\beta = 0.23, p_{Holm} < .01$), Liberal+ opinion ($\beta = 0.18, p_{Holm} < .01$), and budget allocation ($\beta = 2.88, p_{Holm} < .01$), and was additionally the strongest predictor of post-task AI Trust ($\beta = 0.37, p_{Holm} < .001$). AI Education was not a significant predictor of any persuasion outcome after correction (all $p_{Holm} > .10$), but did significantly increase post-task AI Trust ($\beta = 0.20, p_{Holm} < .05$). As before, condition and party coefficients remained substantively consistent with our primary models.

This divergence, greater self-assessed knowledge predicting resistance to persuasion while more frequent everyday use predicts greater persuasion and trust, suggests these two measures may capture distinct constructs rather than a single underlying "AI familiarity" dimension.

5.5 Correlation of Post-Task AI Trust and Strength of Persuasion of Partisan Biased AI

Our analysis of user characteristics revealed an inconsistent relationship between post-task AI trust and persuasion, motivating a closer examination of this association. To explore this, we added post-task AI Trust as a covariate to the H1 and H2 models, estimated separately by party. AI Trust was tested two-tailed, as we had not pre-registered a direction; `Risk`, `Causal`, and their interaction remained one-tailed as in H1 and H2. The results showed that higher post-task AI trust was significantly associated with greater persuasion across all three outcomes for both Democrats and Republicans (all $p < .05$; full regression results are reported in Section C.6). Adjusting for AI trust did not substantially change the estimated effects of `Risk` or `Causal` relative to the H1 and H2 models.

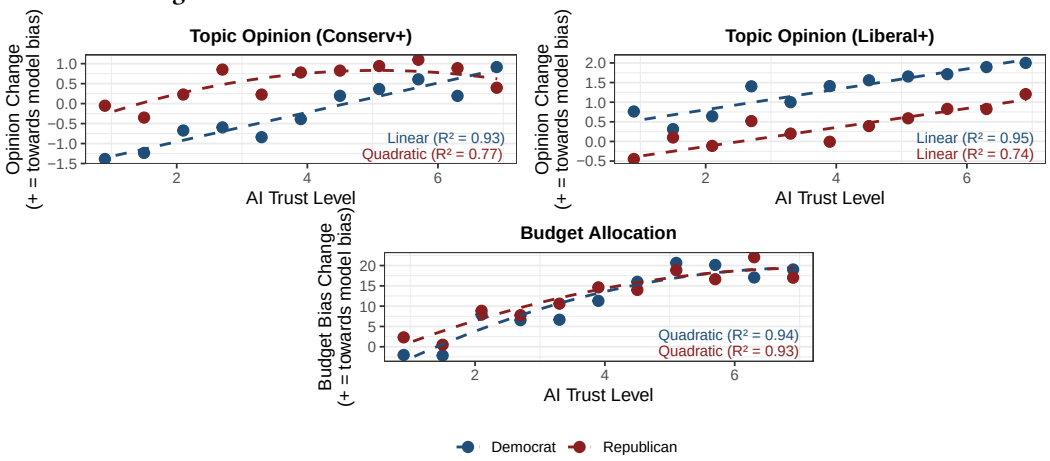
Additionally, Figure 4 illustrates these relationships by plotting mean change in opinion and budget allocation across levels of post-task AI trust, with separate fitted trends (either linear or quadratic based on best fit) for Democrats and Republicans. Across outcomes, AI trust was strongly associated with the magnitude of persuasion, with fitted models explaining a large proportion of the variation (average $R^2 = 0.88$). The relationship was generally positive, although some outcomes showed nonlinear patterns, including an inverted-U relationship for Republicans' Conserv+ opinion and a ceiling effect for budget allocation at the highest levels of trust. Republicans also generally reported higher post-task AI trust than Democrats, particularly for the Conserv+ topic and budget allocation tasks.

Table 8: AI Familiarity Measures as Predictors of Persuasion and AI Trust (Pooled Model)

	Conserv+ Opinion	Liberal+ Opinion	Budget Allocation	AI Trust
Predictor	*b* (SE)	*b* (SE)	*b* (SE)	*b* (SE)
Risk	−0.13(0.13)	0.01(0.13)	**-3.04****(1.49)	0.05(0.11)
Causal	**-0.25****(0.13)	0.05(0.12)	0.00(1.54)	0.02(0.11)
Risk × Causal	**0.38****(0.18)	−0.03(0.18)	2.24(2.18)	−0.04(0.16)
Republican	**1.01****(0.09)	**-0.56****(0.10)	**4.94****(1.14)	**0.81****(0.08)
AI Knowledge	**-0.18****(0.07)	**-0.23****(0.07)	**-3.51****(0.90)	-0.14*(0.06)
AI Education	0.02(0.08)	0.06(0.07)	0.16(1.06)	**0.20****(0.07)
AI Use	**0.23****(0.06)	**0.18****(0.05)	**2.88****(0.69)	**0.37****(0.05)

Note. Entries are OLS regression coefficients with HC3 robust standard errors in parentheses. Models are estimated on the pooled sample, controlling for party, disclaimer condition (Risk, Causal, Risk×Causal), and the corresponding pre-measure (opinion and budget models only). ** = $p_{\text{Holm}} < .05$ or $p < 0.05$ for disclaimer condition or Republican; * = $p_{\text{raw}} < .05$.

Figure 4: Correlation of Post-Task AI Trust and Persuasion Outcomes



Note: Mean opinion change across binned AI trust 10-levels by party identification. Points represent binned means across 10 equal-width intervals of AI trust. Dashed lines show OLS fits on binned means; R^2 values reflect variance explained in the binned series. Democrats are shown in dark blue, Republicans in dark red..

6 Discussion

This study examined whether model disclaimers can reduce the persuasive influence of politically biased LLMs, and if so, for whom and under what conditions. Our findings suggest that disclaimer effectiveness is conditional, as some disclaimer designs reduced persuasion, but effects varied across users and decision-making contexts and decayed over time. More generally, user characteristics, particularly political affiliation and certain dimensions of AI familiarity, were more consistent predictors of susceptibility to persuasion than disclaimer wording. Below, we present these findings together with the mechanisms that plausibly explain them, followed by implications for transparency intervention design and future research.

6.1 Key Findings: Disclaimers Can Help, but Not Uniformly

Our **first finding is that model disclaimers can reduce persuasion, but their effects are modest and context-dependent**. Relative to the generic/no-causal control, specific and causal disclaimers each reduced political persuasion more in at least one task, but no disclaimer type consistently reduced persuasion across all users or tasks (Section 4.1). Among Democrats, causal disclaimers reduced opinion shifts toward the AI's position in the Topic Opinion Task, but not the Budget Allocation Task. Among Republicans, specific disclaimers reduced shifts toward the AI's preferred priorities in the Budget Allocation Task, but no disclaimer significantly reduced opinion change in the Topic Opinion Task. We similarly found limited evidence that any disclaimer conditional alone reduced post-task AI trust, however an additive effect was seen in

Republicans (Section 4.2) compared to the generic/no-causal control.

These findings are broadly consistent with the one existing field test of disclosing an LLM's political bias to users pre-use, which paired the disclosure with source credibility attacks of the model builder (i.e. information about OpenAI's founder, Sam Altman) and found a 28% reduction in persuasion on economic misinformation tasks [23]. Our findings extend this work by isolating disclaimer content from source credibility attacks and examining whether different types of disclaimer information affect downstream persuasion. Further, while previous tasks involved reducing persuasion from misinformation, our tasks were designed to mitigate persuasion from a certain biased framing. These results suggest that model disclaimers can provide a useful but limited form of mitigation.

Our **second finding is that user characteristics, particularly political affiliation, were more consistent predictors of persuasion than disclaimer conditions.** Our results also suggest that this gap is better explained by differential credibility assessment than by inattention. Holding disclaimer conditions constant, Republicans mostly shifted further toward the AI's position and reported higher post-task trust than Democrats across disclaimer conditions (Section 5.1). These partisan differences were larger in magnitude and more consistently significant than the individual disclaimer effects we observed in Finding 1.

One possible explanation is that Republicans simply engaged less with the disclaimer itself. However, our data provide little support for this, as Republicans recalled their assigned disclaimer at rates comparable to Democrats (Section 5.1), and accurate recall was associated with reduced post-task trust in both groups (Section 5.2). This indicates that both groups attended to the disclaimer at similar rates and that this attention results in some consistent downstream opinion change (i.e. reduced AI trust). However, attending to a disclaimer does not necessarily mean finding it credible, and a more likely explanation is that users differed in how they evaluated the disclaimer's credibility once they had registered it. Prior work has shown that Republicans express lower trust than Democrats in institutional sources such as scientists, journalists, and academics [9, 54, 59], and Jia and Lee [49] found that Republicans were more skeptical of fact-checking labels on social media regardless of their content. If users interpreted the AI disclaimer as a claim from the model's developer, or more broadly as a form of fact-checking, the same warning could carry different weight across political groups. We did not directly measure perceived disclaimer credibility, so this explanation remains speculative, but it is consistent with a broader pattern in which prior beliefs and expectations shape how people interpret and respond to new information [61, 69, 81, 111].

Our **third finding is that "AI familiarity" is not a single construct; its components are dissociable in terms of susceptibility to AI persuasion.** We found that greater AI use, measured by frequency of use in a week, predicted both higher post-task trust and greater persuasion, the pathway one would expect if trust drives reliance (Section 5.4). In contrast, self-reported overall AI knowledge predicted lower persuasion but showed no association with post-task trust. Unlike the other two dimensions, formal AI education was not significantly related to either persuasion, but interestingly followed the same pattern as AI use with higher education resulting in higher post-task AI trust.

One possible explanation is that these measures reflect distinct psychological constructs, consistent with leading theories of AI literacy [79]. For example, AI knowledge may reflect evaluative understanding: the ability to critically assess AI-generated information and recognize potential errors or bias. In contrast, frequent AI use may reflect habitual exposure or routine reliance, increasing familiarity and confidence without necessarily improving users' ability to critically evaluate AI-generated content [97, 117]. Formal AI education appears to capture yet another dimension that, at least in our study, was not significantly related to persuasion but was associated with increased trust. This pattern may reflect a "literacy paradox," consistent with recent evidence that greater AI literacy does not necessarily translate into greater resistance to AI influence and may, in some contexts, amplify biases in evaluations of AI-generated content [55]. Although these interpretations remain to be validated, they challenge the common assumption that greater "AI familiarity" or "AI literacy" uniformly increases or decreases susceptibility to AI-generated response. Instead, our findings suggest that different forms of AI familiarity may have qualitatively different and sometimes opposing relationships with trust and persuasion. This dissociation is the first of several cases in which trust and persuasion diverge.

Our **fourth finding is that lower trust was not universally equivalent to lower persuasion.** Trust and reliance are conceptually distinct in the automation literature, with prior research showing that people can trust a system without acting on its output and can act on a system's output for reasons other than trust [64]. Our results mirror this distinction. In Section 5.5 we found that post-task trust in the AI system was strongly associated with persuasion overall (average $R^2 = 0.88$), consistent with prior work linking trust to reliance on AI systems [15, 64]. However, this relationship was not uniform across users, and we observed the dissociation in both directions.

In one direction, trust moved without persuasion following it. Participants from both political parties who correctly recalled the model disclaimer reported lower post-task trust than those who did not accurately recall it, yet only Democrats showed a corresponding reduction in persuasion (Section 5.2). Similarly, Agreeableness and Extraversion were associated with higher post-task trust (Section 5.3), yet only Agreeableness significantly predicted higher persuasion after correcting for multiple comparisons. In the other direction, persuasion moved without trust following it. Including Emotional Stability reducing persuasion with no change in post-task AI trust. Additionally, as noted above, AI knowledge predicted lower persuasion across all tasks with only minimal corresponding association with trust. Together, these two patterns, trust without behavioral change and behavioral change without trust, make a stronger case that trust and reliance are genuinely separable responses to AI-generated information rather than trust simply failing to translate downstream in some cases.

This dissociation echoes a pattern emerging in adjacent AI-disclosure research, which has shown that disclaimers can shift users' perceptions of a system without necessarily changing their downstream judgments or behavior [36, 47, 123]. It may also help explain why the partisan gap in Finding 2 is as large as it is, as prior work suggests that conservatives place greater weight on perceived authorities and established sources of judgment [14, 41],

consistent with Republicans continuing to act on the AI's output even as their reported trust in it declined. Therefore, although trust is an important correlate of persuasion, our findings suggest that reducing trust alone may be insufficient to reduce the persuasive influence of biased AI.

Finally, our fifth finding is that disclaimer effects attenuated over time and were no longer statistically detectable one week later. Although some disclaimer conditions reduced persuasion during the initial interaction relative to the control, we found no evidence that these effects remained statistically significant at the one-week follow-up (Section 4.3). In several cases, however, the follow-up coefficient remained in the same direction as the initial effect but was smaller in magnitude and no longer significant. These patterns are consistent with disclaimer effects fading over time rather than reversing, as a recent study of AI-mediated persuasion on transgender rights similarly found significant short-term attitude change followed by substantial attenuation at a one-week follow-up [19]. These findings suggest that brief interactions with AI may produce measurable shifts in attitudes that are difficult to sustain over time. Thus, model disclaimers may interrupt persuasion during an immediate interaction without producing durable changes in users' subsequent judgments. However, our smaller follow-up sample limits our ability to distinguish genuine attenuation from a loss of statistical power. Further, we cannot discount the possibility that disclosure of the experiment after the main study affected participants' opinions at follow-up, as this disclosure occurred after the initial experiment but before the follow-up survey.

Overall, our findings support the claim that **disclaimer effectiveness is conditional rather than universal**. Improving transparency may therefore require attention not only to what a disclaimer says but also to who delivers it, how it is presented, where it is deployed, and whether disclosure is sufficient for the decision at hand. We use these findings as a foundation for thinking more broadly about how transparency interventions can be designed and evaluated to reduce the influence of biased AI.

6.2 Implications for Transparency Intervention Design

Our findings suggest three implications for interventions intended to reduce the influence of biased AI. First, disclaimer design can be improved by attending to its content, source, and presentation. Second, the conditional nature of our findings highlights the importance of understanding where interventions work and where their effects may not generalize. Finally, reducing AI influence may require going beyond disclaimers, instead prioritizing interventions that change how people interact with AI through improving critical evaluation of its outputs.

6.2.1 Disclaimer Design. Our findings provide several concrete directions for improving disclaimer design itself. We organize these implications around three features of a disclaimer: its content, its source, and its presentation.

Content. Our first finding showed that specific and causal disclaimers, which conveyed concrete information about how and why the model might be biased, outperformed the generic disclaimer in reducing persuasion, even though neither did so universally (Section 4.1). These results showed that disclaimer content, specifically risk specification, can shape its effectiveness. Given this result, one possible design approach is to develop standardized, quantified disclosure formats analogous to nutrition labels or financial risk disclosures. In privacy research, Kelley et al. [2009, 2010] proposed a similar "nutrition label" approach that organizes complex information into a standardized format, making it easier for users to locate and interpret relevant characteristics of a system. Their subsequent evaluation found that standardized presentations improved the speed and accuracy with which users could find information in privacy policies [53]. Similarly, such practices have been suggested for political bias in LLMs [29].

An analogous approach could be applied to AI-generated content. Rather than relying on a generic statement that "AI can be biased," a standardized AI disclosure could report the type of risk being disclosed, the evidence supporting the disclosure, and, when possible, substantive citations for quantitative information about the model's observed behavior. For example, a system could indicate that it has exhibited asymmetric political framing on a defined evaluation set, provide the magnitude of that asymmetry, and distinguish between evidence about the model's behavior and claims about the causes of that behavior. Such a format could make disclosures more comparable across systems while reducing the burden on users to interpret heterogeneous warning language.

Source. Our second findings suggest that users may attend to a disclaimer without necessarily treating it as equally credible, raising the possibility that disclosures originating solely from AI developers may have limited persuasive authority for some users (Section 5.1, Section 5.2). One potential response is to incorporate independent third parties into the disclosure process. A growing ecosystem of organizations now evaluates AI systems independently of their developers, including organizations such as METR, Apollo Research, and the AI Evaluator Forum, alongside academic and nonprofit evaluators [1, 7, 73]. Additionally, industry efforts increasingly recognize this role as well, as Anthropic has called for a broader third-party evaluation ecosystem [6], and OpenAI has described incorporating external assessors into model evaluations [86]. This emerging infrastructure could make disclaimers more credible by separating the production of the system from the evaluation of its risks, analogous to the role that independent auditors and certification bodies play in other domains. More broadly, improving AI disclosures may require not only better warning language but also an institutional ecosystem capable of producing trusted information about AI systems.

Presentation Our fifth finding that disclaimer effects diminished one week later is consistent with fading memory and can indicate that these interventions might require greater effect at presentation. Security-warning research using brain fMRI imaging has shown that repeated exposure to an identical warning produces measurable neurological habituation and that varying the warning's presentation, not its underlying message, restores attentional response and downstream effectiveness [3]. If a similar process contributes to the fading effects found in Section 4.3, the fix may not be a more persistent or more frequent identical warning but a polymorphic one, meaning rotating the disclaimer's framing, visual presentation, or specific wording across repeated exposures within

a single deployment. The goal would not be to make warnings more visually salient for its own sake but to preserve users' attention to information that becomes increasingly easy to ignore through repetition. This possibility is particularly relevant as AI systems move from occasional interactions toward persistent workplace and consumer tools, where a warning encountered once may become a familiar piece of interface furniture rather than an active input into decision-making.

6.2.2 Intervention Evaluation. These design refinements all assume that a disclaimer, however well constructed, is being deployed into the right context and to the right user. Our first finding is also a caution on this point, as no disclaimer type we tested worked uniformly across users or tasks. This heterogeneity should not necessarily be interpreted as evidence that disclaimer interventions are fundamentally flawed. Instead, it suggests that understanding an intervention's limits is itself an important component of intervention design.

This distinction matters because there is a risk of treating an intervention's average effect as evidence that it can be deployed broadly. An intervention that reduces persuasion in one population or decision context may have little effect in another, even when the underlying warning is identical. Our findings therefore suggest that interventions should be validated across the populations, tasks, and outcomes in which they are intended to be deployed. For AI disclaimers, this means testing not only whether a warning changes immediate judgments in a laboratory task but also whether it changes reliance, whether those effects persist, and whether users with different prior beliefs respond differently.

Importantly, this principle is not equivalent to arguing that every context requires a bespoke intervention. Rather, it suggests that intervention designers should know the empirical boundary conditions of a given intervention before assuming that it will generalize. In this sense, heterogeneity is not necessarily a defect to be eliminated. It can instead be useful information about when an intervention is likely to succeed and when other forms of mitigation are needed.

6.2.3 Beyond Disclaimers. Our findings suggest that disclaimers should be viewed as one component of a broader intervention toolkit rather than as a complete solution to AI-mediated persuasion. In particular, our fourth finding that lower trust does not reliably translate into lower persuasion suggests that changing users' perceptions of AI may be insufficient, regardless of how a disclaimer is worded, sourced, or presented. This suggests targeting either the decision process itself or users' ability to evaluate AI outputs.

One approach is to intervene directly in the decision process rather than users' general perceptions of AI. For example, cognitive forcing functions introduce friction [80], requiring users to actively engage with AI recommendations before accepting them, such as by clicking an "I understand" button, retyping a message, or completing a checklist. Buçinca et al. [13] found that such interventions reduced AI overreliance relative to simpler explainable-AI approaches, although effects varied across users.

Individual-level versions of cognitive forcing functions are already used for other high-stakes, hard-to-reverse actions. For instance, GitHub[4], a widely used platform for hosting and managing software code, requires users to retype a repository's name before changing its visibility. An analogous pattern for a politically consequential AI recommendation might require users to briefly articulate their own reasoning or actively confirm that they have considered an alternative view before proceeding. Cognitive forcing can also operate at the institutional level. Aviation checklists, clinical decision protocols, and financial compliance procedures formalize required steps before a consequential decision is finalized, regardless of how confident or rushed the decision-maker feels at the moment. An analogous safeguard for AI-assisted decisions might require that users independently identify the relevant evidence, articulate their initial judgment, and document their reasons for accepting or rejecting an AI recommendation.

AI literacy training is the second and more durable direction. Our third finding, that AI use and AI knowledge predicted trust and persuasion in opposite ways, suggests that increasing users' comfort and familiarity with AI systems may not produce the evaluative skills needed to resist AI persuasion. This distinction is particularly relevant as governments and employers invest in large-scale AI upskilling programs. For example, the UK's AI Skills Boost program aims to provide AI training to 10 million workers by 2030, with foundational training emphasizing practical skills such as writing instructions for AI systems, using AI to support routine tasks, and understanding risks associated with AI use [112]. The U.S. Department of Labor has similarly launched its Make America AI-Ready course [113], which emphasizes practical AI skills and use as well. Our findings suggest that programs like these may need to go beyond building AI fluency to explicitly develop the evaluative skills needed to recognize and resist AI influence.

These directions suggest that refining disclaimer wording is only one part of a larger design space. Validating interventions for their actual context of use, and complementing informational disclosure with structural and educational interventions may do more to reduce the persuasive influence of biased AI systems than refining disclaimer wording alone.

6.3 Future Work

We identify several open questions raised by our findings that could warrant direct empirical attention. First, our explanation for the partisan gap in Finding 2, that Republicans and Democrats differ in the credibility they assign to the same disclaimer, was not measured directly in this study. Future work could assess perceived disclaimer credibility, for example by manipulating disclaimer source or messenger identity, to test whether it helps explain the persuasion gap we observed. Second, our data does not explain why particular disclaimer types were effective for particular tasks and parties. We encourage future tests of whether task characteristics, such as the salience of partisan identity in a given decision, systematically moderate the effectiveness of different disclaimer types. Finally, the design directions proposed in Section 6.2, including standardized disclosure formats, third-party sourcing, cognitive forcing functions, and adversarial AI literacy training, are motivated by our

[4]https://github.com

findings and adjacent literature rather than interventions tested in this study. These approaches could provide promising avenues for reducing AI influence, but their effectiveness remains an empirical question that could be evaluated, ideally within the paradigm used here.

6.4 Limitations

This work has several limitations that should inform the interpretation of our findings. First, many of the analyses discussed in this paper were exploratory and correlational, particularly those examining political identity, personality, AI use, AI knowledge, and AI education. As such, these results should be interpreted as hypothesis-generating rather than providing evidence of causal relationships. Second, the generalizability of our findings is limited because the study focused on U.S. political affiliations and should be replicated in other political and cultural contexts. Third, our use of online recruitment introduces the limitations associated with non-representative samples. Fourth, we evaluated a single language model, GPT-4o-mini [84], using one prompting strategy, limiting the extent to which our findings generalize to other public LLMs or different forms and degrees of political bias. Lastly, we include only two types of tasks in our experimentation, which although diverse, do not span the whole space of political decision-making. Future work could test whether these effects generalize to other salient political tasks and contexts.

6.5 Conclusion

Our findings suggest that improving disclaimer effectiveness may require moving beyond one-size-fits-all disclaimers. Disclaimers can reduce persuasion, but their effects depend on users and decision contexts and may not persist over time. More broadly, the divergence between trust and persuasion highlights the limits of interventions focused solely on changing how users perceive AI. Mitigating AI influence may therefore require a broader approach that combines better disclosures with cognitive safeguards and AI literacy that strengthens users' ability to critically evaluate AI outputs. As AI systems become increasingly embedded in consequential decision-making, effective safeguards may therefore require not a single solution, but a broader ecosystem of interventions that helps people recognize, question, and appropriately respond to AI influence.

Ethics and Privacy Statement

This study demonstrates that politically biased LLM outputs can influence users and that current model disclaimers provide only partial and potentially temporary protection against these persuasive effects. Our findings on differential user susceptibility and the methods used to study model bias could potentially be misapplied to better understand how AI systems influence particular users; however, we present them to support the development and evaluation of more effective safeguards against unwanted AI influence. Additionally, participants were not informed of the model's bias during the interaction to preserve ecological validity, however, the study was deemed exempt by an IRB, and all participants were subsequently debriefed about the model's bias and given the opportunity to withdraw their data.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS2554007 This research was developed with funding from the Defense Advanced Research Projects Agency's (DARPA) SciFy program (Agreement No. HR00112520300). The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. We gratefully acknowledge gift funding from Coefficient Giving. This research was also supported in part by DARPA under the ITM program (FA8650-23-C-7316) and NSF grant No. 2230466.

References

[1] AI Evaluator Forum. 2026. AI Evaluator Forum. https://aievaluatorforum.org/. Accessed: 2026-08-23.

[2] American Postal Workers Union. 2025. *Postal Banking.* Accessed: 2025-05-23. https://apwu.org/postal-banking

[3] Bonnie Brinton Anderson, C. Brock Kirwan, Jeffrey L. Jenkins, David Eargle, Seth Howard, and Anthony Vance. 2015. How Polymorphic Warnings Reduce Habituation in the Brain: Insights from an fMRI Study. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015), 2883–2892. doi:10.1145/2702123.2702322

[4] ANSI Committee Z535. 1998. American National Standard: Safety Color Code, ANSI Z535.1-1998. National Electrical Manufacturers Association, Rosslyn, VA.

[5] Anthropic. 2026. *Claude Mythos Preview System Card.* System Card. Anthropic. Version published April 7, 2026. https://www-cdn.anthropic.com/7624816413e9b4d2e3ba620c5a5e091b98b190a5/Claude%20Mythos%20Preview%20System%20Card.pdf

[6] Anthropic. 2026. A new initiative for developing third-party model evaluations. https://www.anthropic.com/news/a-new-initiative-for-developing-third-party-model-evaluations. Accessed: 2026-08-23.

[7] APOLLO Research. 2026. APOLLO Research. https://www.apolloresearch.ai/. Accessed: 2026-08-23.

[8] Peter M. Aronow, Jonathon Baron, and Lauren Pinson. 2019. A Note on Dropping Experimental Subjects who Fail a Manipulation Check. *Political Analysis* 27, 4 (2019), 572–589. doi:10.1017/pan.2019.5

[9] Flávio Azevedo and John T. Jost. 2021. The ideological basis of antiscientific attitudes: Effects of authoritarianism, conservatism, religiosity, social dominance, and system justification. *Group Processes & Intergroup Relations* 24, 4 (June 2021), 518–549. doi:10.1177/1368430221990104

[10] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623.

[11] Jeffrey M. Berry and David F. Arons. 2005. *A Voice for Nonprofits.* Brookings Institution Press. http://www.jstor.org/stable/10.7864/j.ctv10h9f03

[12] Curt C. Braun, Paul B. Mine, and N. Clayton Silver. 1995. The influence of color on warning label perceptions. *International Journal of Industrial Ergonomics* 15, 3 (1995), 179–187. doi:10.1016/0169-8141(94)00036-3

[13] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5 (April 2021), 1–21. Issue CSCW1. doi:10.1145/3449287

[14] Dana R. Carney, John T. Jost, Samuel D. Gosling, and Jeff Potter. 2008. The Secret Lives of Liberals and Conservatives: Personality Profiles, Interaction Styles, and the Things They Leave Behind. *Political Psychology* 29, 6 (2008), 807–840. https://www.jstor.org/stable/20447169

[15] Erin K. Chiou and John D. Lee. 2023. Trusting automation: Designing for responsivity and resilience. *Human Factors* 65, 1 (2023), 137–165. doi:10.1177/00187208211009995

[16] Cindy K. Chung and James W. Pennebaker. 2007. The Psychological Functions of Function Words. *Social Communication* (2007), 343–359.

[17] Katherine Clayton, Spencer Blair, Jonathan A. Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, Morgan Sandhu, Rachel Sang, Rachel Scholz-Bright, Austin T. Welch, Andrew G. Wolff, Amanda Zhou, and Brendan Nyhan. 2020. Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media. *Political Behavior* 42, 4 (2020), 1073–1095. doi:10.1007/s11109-019-09533-0

[18] Donald Cohen. 2021. *Postal Banking Is Finally a Reality in (Some of) the United States.* Accessed: 2025-05-23. https://inthesetimes.com/article/postal-banking-

post-office-apwu

[19] Charles Crabtree, John B Holbein, Mitchell Bosley, and Semra Sevi. 2026. Can AI help reduce prejudice? Evaluating the effectiveness of AI-powered personalized persuasion on support for transgender rights. *PNAS Nexus* 5, 7 (07 2026), pgag066. doi:10.1093/pnasnexus/pgag066

[20] Shannon M. Cruz and Christopher J. Carpenter. 2024. The Roles of Identity- and Belief-Driven Motivated Reasoning and Source Credibility in Persuasion on Climate Change Policy. *Journal of Language and Social Psychology* 43, 5 (2024), 592–619. doi:10.1177/0261927X241291572

[21] Jelena Cupać and Mitja Sienknecht. 2024. Regulate against the machine: how the EU mitigates AI harm to democracy. *Democratization* 31, 5 (2024), 1067–1090. doi:10.1080/13510347.2024.2353706

[22] R. E. Dewar. 1999. Design and Evaluation of Public Information Symbols. In *Visual Information for Everyday Use: Design and Research Perspectives*, H. J. G. Zwaga, T. Boersma, and H. C. M. Hoonhout (Eds.). Taylor and Francis, London, 111–117.

[23] M. DiGiuseppe and J. Robison. 2026. Perceived political bias in large language models reduces persuasive abilities. pgag324 pages. doi:10.1093/pnasnexus/pgag324

[24] Peter H. Ditto, Jared B. Celniker, Shiri Spitz Siddiqi, Mertcan Güngör, and Daniel P. Relihan. 2025. Partisan Bias in Political Judgment. *Annual Review of Psychology* 76, 1 (Jan. 2025), 717–740. doi:10.1146/annurev-psych-030424-122723

[25] Emily DiVito. 2022. *Banking for All: How the USPS Could Provide Public Banking.* Accessed: 2025-05-23. https://rooseveltinstitute.org/blog/banking-for-all/

[26] Ullrich K. H. Ecker, Stephan Lewandowsky, and David T. W. Tang. 2010. Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition* 38, 8 (2010), 1087–1100. doi:10.3758/MC.38.8.1087

[27] Mats Faulborn, Indira Sen, Max Pellert, Andreas Spitz, and David Garcia. 2025. Only a Little to the Left: A Theory-grounded Measure of Political Bias in Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 31684–31704. doi:10.18653/v1/2025.acl-long.1529

[28] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 11737–11762.

[29] Jillian Fisher, Ruth E. Appel, Chan Young Park, Yujin Potter, Liwei Jiang, Taylor Sorensen, Shangbin Feng, Yulia Tsvetkov, Margaret E. Roberts, Jennifer Pan, Dawn Song, and Yejin Choi. 2025. Position: political neutrality in AI is impossible — but here is how to approximate it. In *Proceedings of the 42nd International Conference on Machine Learning (ICML'25)*. JMLR.org, Vancouver, Canada, Article 3273, 60 pages.

[30] Jillian Fisher, Shangbin Feng, Robert Aron, Thomas Richardson, Yejin Choi, Daniel W Fisher, Jennifer Pan, Yulia Tsvetkov, and Katharina Reinecke. 2025. Biased LLMs can Influence Political Decision-Making. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 6559–6607. doi:10.18653/v1/2025.acl-long.328

[31] Chloë FitzGerald and Samia Hurst. 2017. Implicit Bias in Healthcare Professionals: A Systematic Review. *BMC Medical Ethics* 18, 1 (March 2017), 19. doi:10.1186/s12910-017-0179-8

[32] The Heritage Foundation. 2020. *Why Postal Banking Would Help Neither the Poor Nor the Postal Service.* Accessed: 2025-05-23. https://www.heritage.org/monetary-policy/commentary/why-postal-banking-would-help-neither-the-poor-nor-the-postal-service

[33] J. P. Frantz, T. P. Rhoades, S. L. Young, and J. A. Schiller. 2000. Assessing the Effects of Adding Messages to Warning Labels. In *Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting*, Vol. 4. Human Factors and Ergonomics Society, Santa Monica, CA, 818–821.

[34] Marian Friestad and Peter Wright. 1994. The Persuasion Knowledge Model: How People Cope with Persuasion Attempts. *Journal of Consumer Research* 21, 1 (1994), 1–31. https://www.jstor.org/stable/2489738

[35] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* 50, 3 (Sept. 2024), 1097–1179. doi:10.1162/coli_a_00524

[36] Dilrukshi Gamage, Dilki Sewwandi, Min Zhang, and Arosha K Bandara. 2025. Labeling Synthetic Content: User Perceptions of Label Designs for AI-Generated Content on Social Media. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (2025-04-25) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–29. doi:10.1145/3706598.3713171

[37] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. arXiv:2209.07858 [cs.CL] https://arxiv.org/abs/2209.07858

[38] Alan S. Gerber, Gregory A. Huber, David Doherty, Conor M. Dowling, and Costas Panagopoulos. 2013. Big Five personality traits and responses to persuasive appeals: Results from voter turnout experiments. *Political Behavior* 35, 4 (2013), 687–728. doi:10.1007/s11109-012-9216-y

[39] Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality* 37, 6 (2003), 504–528. doi:10.1016/S0092-6566(03)00046-1

[40] Rebecca Hofstein Grady, Peter H. Ditto, and Elizabeth F. Loftus. 2021. Nevertheless, partisanship persisted: fake news warnings help briefly, but bias returns with time. *Cognitive Research: Principles and Implications* 6 (2021), 52. doi:10.1186/s41235-021-00315-z

[41] Jesse Graham, Jonathan Haidt, and Brian A. Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology* 96, 5 (2009), 1029–1046. doi:10.1037/a0015141

[42] Robert Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill, New York.

[43] John Heltman. 2020. Postal Banking is a Solution, but to Which Problem? Accessed: 2025-05-23. https://www.americanbanker.com/podcast/postal-banking-is-a-solution-but-to-which-problem

[44] Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. 2024. Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. *ACM Journal on Responsible Computing* 1, 2 (2024), 11:1–11:52. doi:10.1145/3631326

[45] K. Hu, E. De Rosa, and A. K. Anderson. 2020. Yellow is for safety: perceptual and affective perspectives. *Psychological Research* 84, 7 (Oct. 2020), 1912–1919. Epub 2019 Apr 29. doi:10.1007/s00426-019-01186-2

[46] Howard Husock. 2023. *Removing the Tax Exemption for Religious Congregations Is the Wrong Move for Our Polarized Nation.* Accessed: 2025-05-24. https://www.aei.org/op-eds/removing-the-tax-exemption-for-religious-congregations-is-the-wrong-move-for-our-polarized-nation/

[47] Lujain Ibrahim, Myra Cheng, Cinoo Lee, Pranav Khadpe, Desmond Ong, Dan Jurafsky, and Diyi Yang. 2026. Warning labels shift perceptions of sycophantic AI, but not its influence. arXiv:2606.21317 [cs.HC] doi:10.48550/arXiv.2606.21317

[48] Westwood Sean J., Grinner Justin, and Hall Andrew B. 2026. *Measuring Perceived Slant in Large Language Models Through User Evaluations.* https://www.gsb.stanford.edu/faculty-research/working-papers/measuring-perceived-slant-large-language-models-through-user

[49] Chenyan Jia and Taeyoung Lee. 2024. Journalistic interventions matter: Understanding how Americans perceive fact-checking labels. *Harvard Kennedy School Misinformation Review* (April 2024). doi:10.37016/mr-2020-138

[50] Oliver P. John and Sanjay Srivastava. 1999. The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research* 2 (1999), 102–138. https://psycnet.apa.org/record/1999-04371-004

[51] John T. Jost. 2021. *Left and right: The psychological significance of a political distinction.* Oxford University Press, New York, NY, US. Pages: xviii, 392.

[52] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. 2009. A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security* (Mountain View, California, USA) *(SOUPS '09)*. Association for Computing Machinery, New York, NY, USA, Article 4, 12 pages. doi:10.1145/1572532.1572538

[53] Patrick Gage Kelley, Lucian Cesca, Joanna Bresee, and Lorrie Faith Cranor. 2010. Standardizing privacy notices: an online study of the nutrition label approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1573–1582. doi:10.1145/1753326.1753561

[54] Brian Kennedy and Emma Kikuchi. 2026. *Americans' confidence in scientists.* Section: Trust in Science. https://www.pewresearch.org/science/2026/01/15/americans-confidence-in-scientists/

[55] WooJin Kim and Yuhosua Ryoo. 2026. The literacy paradox: How AI literacy amplifies biases in evaluating AI-generated news articles. *International Journal of Information Management* 86 (2026), 102992. doi:10.1016/j.ijinfomgt.2025.102992

[56] J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.* Technical Report Research Branch Report 8-75. Naval Technical Training Command, Millington TN Research Branch.

[57] D. J. Koehler. 1991. Explanation, imagination, and confidence in judgment. *Psychological Bulletin* 110, 3 (1991), 499–519. doi:10.1037/0033-2909.110.3.499

[58] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. https://arxiv.org/abs/2506.08872

[59] Patrick W. Kraft, Milton Lodge, and Charles S. Taber. 2015. Why people "don't trust the evidence": Motivated reasoning and scientific beliefs. *Annals of the American Academy of Political and Social Science* 658, 1 (2015), 121–133. doi:10.1177/0002716214554758

[60] Anne C Kroon, Toni G L A van der Meer, and Thomas Pronk. 2022. Does Information about Bias Attenuate Selective Exposure? The Effects of Implicit Bias Feedback on the Selection of Outgroup-Rich News. *Human Communication Research* 48, 2 (02 2022), 346–373. arXiv:https://academic.oup.com/hcr/article-pdf/48/2/346/42986977/hqac004.pdf doi:10.1093/hcr/hqac004

[61] Ziva Kunda. 1990. The case for motivated reasoning. *Psychological Bulletin* 108, 3 (1990), 480–498. doi:10.1037/0033-2909.108.3.480

[62] Victor A. F. Lamme. 2003. Why visual attention and awareness are different. *Trends in Cognitive Sciences* 7, 1 (Jan. 2003), 12–18. doi:10.1016/s1364-6613(02)00013-x

[63] David S. Lee. 2009. Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *The Review of Economic Studies* 76, 3 (2009), 1071–1102. http://www.jstor.org/stable/40247633

[64] John D. Lee and Katrina A. See. 2004. Trust in automation: designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80. doi:10.1518/hfes.46.1.50_30392

[65] Thomas J. Leeper and Rune Slothuus. 2014. Political Parties, Motivated Reasoning, and Public Opinion Formation. *Political Psychology* 35 (2014), 129–156. https://www.jstor.org/stable/43783792

[66] Angelica Lermann Henestrosa and Joachim Kimmerle. 2025. "Always check important information!" - The role of disclaimers in the perception of AI-generated content. *Computers in Human Behavior: Artificial Humans* 4 (2025), 100142. doi:10.1016/j.chbah.2025.100142

[67] Hause Lin, Gabriela Czarnek, Benjamin Lewis, Joshua P. White, Adam J. Berinsky, Thomas Costello, Gordon Pennycook, and David G. Rand. 2025. Persuading voters using human–artificial intelligence dialogues. *Nature* 648, 8093 (2025), 394–401. doi:10.1038/s41586-025-09771-9

[68] Melanie G. Long and Steven Pressman and. 2024. Postal banking and US cash transfer programs: a solution to insufficient financial infrastructure? *Review of Social Economy* 82, 2 (2024), 213–240. arXiv:https://doi.org/10.1080/00346764.2023.2259362 doi:10.1080/00346764.2023.2259362

[69] Charles G. Lord, Lee Ross, and Mark R. Lepper. 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology* 37, 11 (1979), 2098–2109. doi:10.1037/0022-3514.37.11.2098

[70] Cameron Martel and David G. Rand. 2023. Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology* 54 (2023), 101710. doi:10.1016/j.copsyc.2023.101710

[71] Lloyd H. Mayer. 2016. Dark Money and Shadow Parties: The Real Problem with Citizens United. *Notre Dame Law Review* 89, 3 (2016), 901–953.

[72] Melanie J. McGrath, Oliver Lack, James Tisch, and Andreas Duenser. 2025. Measuring trust in artificial intelligence: validation of an established scale and its short form. *Frontiers in Artificial Intelligence* 8 (2025). doi:10.3389/frai.2025.1582880

[73] METR. 2026. Model Evaluation & Threat Research. https://metr.org/. Accessed: 2026-08-23.

[74] Luise Metzger, Linda Miller, Martin Baumann, and Johannes Kraus. 2024. Empowering Calibrated (Dis-)Trust in Conversational Agents: A User Study on the Persuasive Power of Limitation Disclaimers vs. Authoritative Style. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, New York, NY, USA, Article 481, 19 pages. doi:10.1145/3613904.3642122

[75] Luise Metzger, Linda Miller, Martin Rudolf Karl Baumann, and Johannes Kraus. 2024. Empowering Calibrated (Dis-)Trust in Conversational Agents: A User Study on the Persuasive Power of Limitation Disclaimers vs. Authoritative Style. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3613904.3642122

[76] Gautam Misra Mitra, Supratik. 2024. *Navigating AI Safety: A Socio-Technical and Risk-based Approach to Policy Design*. https://techpolicy.press/navigating-ai-safety-a-sociotechnical-and-riskbased-approach-to-policy-design

[77] Saif M. Mohammad. 2025. NRC VAD Lexicon v2: Norms for Valence, Arousal, and Dominance for over 55k English Terms. arXiv:2503.23547 [cs.CL] https://arxiv.org/abs/2503.23547

[78] Jacob M. Montgomery, Brendan Nyhan, and Michelle Torres. 2018. How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It. *American Journal of Political Science* 62, 3 (2018), 760–775. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12357 doi:10.1111/ajps.12357

[79] Davy Tsz Kit Ng, Jac Ka Lok Leung, Samuel Kai Wah Chu, and Maggie Shen Qiao. 2021. Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence* 2 (2021), 100041. doi:10.1016/j.caeai.2021.100041

[80] Donald A Norman and Peter Berkrot. 2013. *The design of everyday things*. Vol. 272. Basic books New York.

[81] Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32, 2 (2010), 303–330.

[82] Bureau of Cyberspace and Digital Policy. 2024. *Risk Management Profile for AI and Human Rights*. https://2021-2025.state.gov/risk-management-profile-for-ai-and-human-rights/

[83] OpenAI. 2023. gpt-3.5-turbo-1106. https://platform.openai.com/docs/models/gpt-3-5-turbo. Accessed: 2023-09-02.

[84] OpenAI. 2024. GPT-4o Mini. https://openai.com. Accessed via OpenAI API.

[85] OpenAI. 2026. GPT-5.4 mini: Advancing Cost-Efficient Intelligence. https://developers.openai.com/api/docs/models/gpt-5.4-mini.

[86] OpenAI. 2026. Third-party cyber evaluations involving OpenAI models. https://openai.com/index/third-party-cyber-evaluations-involving-openai-models/. Accessed: 2026-08-23.

[87] Christopher J. Pannucci and E. G. Wilkins. 2010. Identifying and Avoiding Bias in Research. *Plastic and Reconstructive Surgery* 126, 2 (2010), 619–625. doi:10.1097/PRS.0b013e3181de24bc

[88] Paul Arnsberger and Melissa Ludlum and Margaret Riley and Mark Stanton. 2009. *A History of the Tax-Exempt Sector: An SOI Perspective*. Technical Report. Statistics of Income Division, Internal Revenue Service. Accessed: 2025-05-24. https://www.irs.gov/pub/irs-soi/tehistory.pdf

[89] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70 (2017), 153–163. doi:10.1016/j.jesp.2017.01.006

[90] Gordon Pennycook, Adam Bear, Evan T. Collins, and David G. Rand. 2020. The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings. *Management Science* 66, 11 (2020), 4944–4957. https://ideas.repec.org//a/inm/ormnsc/v66y2020i11p4944-4957.html

[91] Richard E. Petty and John T. Cacioppo. 1986. The Elaboration Likelihood Model of Persuasion. In *Advances in Experimental Social Psychology*. Vol. 19. Academic Press, 123–205. doi:10.1016/S0065-2601(08)60214-2

[92] Pew Research Center. 2019. *Views of Government and the Nation*. Accessed: 2025-05-24. https://www.pewresearch.org/politics/2019/12/17/views-of-government-and-the-nation/

[93] Yujin Potter, Shiyang Lai, Junsol Kim, James Evans, and Dawn Song. 2024. Hidden Persuaders: LLMs' Political Leaning and Their Influence on Voters. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 4244–4275. doi:10.18653/v1/2024.emnlp-main.244

[94] Luca Rettenberger, Markus Reischl, and Mark Schutera. 2025. Assessing political bias in large language models. *Journal of Computational Social Science* 8, 2 (2025), 1–17. https://ideas.repec.org//a/spr/jcsosc/v8y2025i2d10.1007_s42001-025-00376-w.html

[95] Nancy Rhodes and Wendy Wood. 1992. Self-Esteem and Intelligence Affect Influenceability: The Mediating Role of Message Reception. *Psychological Bulletin* 111, 1 (1992), 156–171. doi:10.1037/0033-2909.111.1.156

[96] Erich Riesen. 2025. A sociotechnological-system approach to AI ethics. *AI & SOCIETY* 40, 8 (April 2025), 6231–6243. doi:10.1007/s00146-025-02335-5

[97] Evan F. Risko and Sam J. Gilbert. 2016. Cognitive offloading. *Trends in Cognitive Sciences* 20, 9 (2016), 676–688. doi:10.1016/j.tics.2016.07.002

[98] Richard W Robins, Holly M Hendin, and Kali H Trzesniewski. 2001. Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin* 27, 2 (2001), 151–161.

[99] David Rozado. 2024. The political preferences of LLMs. *PLOS ONE* 19, 7 (2024), e0306621. doi:10.1371/journal.pone.0306621

[100] James A. Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178. doi:10.1037/h0077714

[101] Paul Röttger, Musashi Hinck, Valentin Hofmann, Kobi Hackenburg, Valentina Pyatkin, Faeze Brahman, and Dirk Hovy. 2025. IssueBench: Millions of Realistic Prompts for Measuring Issue Bias in LLM Writing Assistance. arXiv:2502.08395 [cs.CL] doi:10.48550/arXiv.2502.08395

[102] Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2024-08), Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 15295–15311. doi:10.18653/v1/2024.acl-long.816

[103] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose Opinions Do Language Models Reflect?. In *Proceedings of the 40th International Conference on Machine Learning* (2023-07-03). PMLR, 29971–30004. https://proceedings.mlr.press/v202/santurkar23a.html

[104] Matthew Shu, Daniel Karell, Keitaro Okura, and Thomas R Davidson. 2026. How latent and prompting biases in AI-generated historical narratives influence opinions. *PNAS Nexus* 5, 3 (2026), pgag022. doi:10.1093/pnasnexus/pgag022

[105] Alyssa H. Sinclair, Matthew L. Stanley, and Paul Seli. 2020. Closed-minded cognition: Right-wing authoritarianism is negatively related to belief updating following prediction error. *Psychonomic Bulletin & Review* 27, 6 (Dec. 2020), 1348–1361. doi:10.3758/s13423-020-01767-y

[106] Danyelle Solomon, Mehrsa Baradaran, and Lily Roberts. 2020. *Creating a Postal Banking System Would Help Address Structural Inequality.* Accessed: 2025-05-23. https://www.americanprogress.org/article/creating-postal-banking-system-help-address-structural-inequality/

[107] Tobias Spinde, Christoph Jeggle, Manuel Haupt, Wolfgang Gaissmaier, and Holger Giese. 2022. How Do We Raise Media Bias Awareness Effectively? Effects of Visualizations to Communicate Bias. *PLoS ONE* 17, 4 (April 2022), e0266204. doi:10.1371/journal.pone.0266204

[108] Jonathan Stray, David Zhai Yang, Steven Luo, Miu Nicole Takagi, and Serina Chang. 2026. Political Neutrality as Balanced Approval: A Large-Scale Human Evaluation of AI Responses. arXiv:2605.28911 [cs.CY] doi:10.48550/arXiv.2605.28911

[109] Natalie Jomini Stroud. 2010. Polarization and Partisan Selective Exposure. *Journal of Communication* 60, 3 (2010), 556–576. doi:10.1111/j.1460-2466.2010.01497.x

[110] Charles S. Taber and Milton Lodge. 2006. Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science* 50, 3 (2006), 755–769. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-5907.2006.00214.x doi:10.1111/j.1540-5907.2006.00214.x

[111] Ben M. Tappin, Gordon Pennycook, and David G. Rand. 2020. Bayesian or biased? Analytic thinking and political belief updating. *Cognition* 204 (2020), 104375. doi:10.1016/j.cognition.2020.104375

[112] UK Government. 2026. AI Skills Boost explainer. https://www.gov.uk/government/publications/ai-skills-boost-explainer. Accessed July 2026.

[113] US Department of Labor. 2026. Make America AI-Ready. https://dol.gov/ai-ready. Free AI literacy text-message course; accessed July 2026.

[114] U.S. Government Accountability Office. 2020. *U.S. Postal Service: Congressional Action is Essential to Enable a Sustainable Business Model.* Technical Report GAO-20-385. U.S. Government Accountability Office. Accessed: 2025-05-23. https://www.gao.gov/assets/gao-20-385.pdf

[115] Monica B. Vela, A. I. Erondu, N. A. Smith, Monica E. Peek, Joan N. Woodruff, and Marshall H. Chin. 2022. Eliminating Explicit and Implicit Biases in Health Care: Evidence and Research Needs. *Annual Review of Public Health* 43 (2022), 477–501. doi:10.1146/annurev-publhealth-052620-103528

[116] Emily K. Vraga. 2015. How Party Affiliation Conditions the Experience of Dissonance and Explains Polarization and Selective Exposure*. *Social Science Quarterly* 96, 2 (2015), 487–502. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/ssqu.12138 doi:10.1111/ssqu.12138

[117] Adrian F. Ward. 2013. Supernormal: How the Internet is changing our memories and our minds. *Psychological Inquiry* 24, 4 (2013), 341–348. doi:10.1080/1047840X.2013.850148

[118] Duane T Wegener and Richard E Petty. 1995. Flexible Correction Processes in Social Judgment: The Role of Naive Theories in Corrections for Perceived Bias. *Journal of Personality and Social Psychology* 68, 1 (1995), 36–51.

[119] Duane T Wegener and Richard E Petty. 1997. The Flexible Correction Model: The Role of Naive Theories of Bias in Bias Correction. In *Advances in Experimental Social Psychology*. Vol. 29. Academic Press, 141–208.

[120] David R. Williams and Lisa A. Cooper. 2019. Reducing Racial Inequities in Health: Using What We Already Know to Take Action. *International Journal of Environmental Research and Public Health* 16, 4 (February 2019), 606. doi:10.3390/ijerph16040606

[121] Michael S. Wogalter, David M. DeJoy, and Kenneth R. Laughery (Eds.). 1999. *Warnings and Risk Communication.* Taylor and Francis, Philadelphia. doi:10.1201/9780203983836

[122] xAI. 2025. *Grok 4 Fast Model Card.* Model Card. xAI. Last updated September 19, 2025. https://data.x.ai/2025-09-19-grok-4-fast-model-card.pdf

[123] Meryl Ye, Robert Kraut, and Steve Rathje. 2026. Observing sycophantic AI validate others reduces its appeal but not its persuasiveness. arXiv:2607.25166 [cs.AI] doi:10.48550/arXiv.2607.25166

[124] S. L. Young and Michael S. Wogalter. 1988. Memory of Instruction Manual Warnings: Effect of Pictorial Icons and Conspicuous Print. In *Proceedings of the Human Factors Society 32nd Annual Meeting*. Human Factors Society, 905–909.

[125] Ceciley (Xinyi) Zhang, Ronald E Rice, and Laurent H Wang. 2026. College students' literacy, ChatGPT activities, educational outcomes, and trust from a digital divide perspective. *New Media & Society* 28, 2 (2026), 673–695. doi:10.1177/14614448241301741

Appendix Table of Contents

A. **Extended Material and Methods**
- Section A.1: Experimental Flow Diagram and Randomization Details
- Section A.2: Power Analysis
- Section A.3: Data Details
- Section A.4: Failed Attention Check Analysis
- Section A.5: Demographics
- Section A.6: Balance Checks - All Participants
- Section A.7: Balance Checks - Excluding Failed Attention Check Participants

B. **Experimental Design Details**
- Section B.1: Model Biasing
- Section B.2: Disclaimer Construction Process
- Section B.3: Disclaimer Presentation
- Section B.4: Current Foundational Disclaimer Examples
- Section B.5: Topic Opinion Task Details
- Section B.6: Budget Allocation Task Details
- Section B.7: Control Variables
- Section B.8: Derived Variables
- Section B.9: Analysis
- Section B.10: Ethical Consideration and IRB
- Section B.11: Consent Form
- Section B.12: Debrief Form
- Section B.13: Qualitative Evaluation Prompt

C. **Additional Results**
- Section C.1: Comparison with Prior Study using No Model Disclaimer
- Section C.2: Qualitative Analysis of Model Skepticism and Distrust
- Section C.3: Pre-Registered Analysis of H4
- Section C.4: Pre-Registered Main and Follow-Up Results Excluding Failed Attention Check Participants
- Section C.5: Pooled Models with Demographic Covariates
- Section C.6: AI Trust and Persuasion Regression Results
- Section C.7: One Week Follow-Up Change Visualization
- Section C.8: Sensitivity Checks

A Extended Materials and Methods

A.1 Experimental Flow Diagram and Randomization Details

See Figure 2 for the full flow of our experiment, and Table 9 for the randomization used and outcomes analyzed. We note that randomization was done per group (within Republican and Democrat participants) independently before the study began.

Randomization	Assignment	Probability
Risk	Inaccuracy/Political Bias	1/2
Causal	Absent/Present	1/2
First Task	Topic Opinion/Budget Allocation	1/2
First Topic Orientation	Conserv+/Liberal+	1/2
Conserv+ Topic	Covenant Marriage/Unilateralism	1/2
Liberal+ Topic	Postal Banking/Nonprofit Tax	1/2

Table 9: Randomization chosen at start of experiment within each partisan group.

A.2 Power Analysis

To determine the appropriate sample size, we drew on data from a previous experiment using a similar design [30]. We conducted a simulation-based power analysis using an ANCOVA-style framework grounded in parameters derived from prior experimental data.

We modeled the post-treatment outcome Y_{post} as a function of the pre-treatment outcome Y_{pre}, our two binary experimental manipulations, `Causal` (C) and `Risk` Type (R), and their interaction:

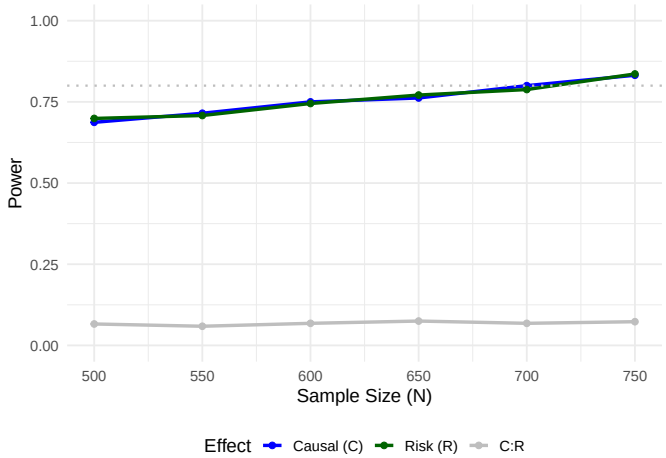


Figure 5: Results of simulated power analysis. We display the estimated power based on $1,000$ samples for each sample size (N), for both main effects (C, R) and the interaction effecct (C:R).

$$Y_{post} = \beta_{pre} Y_{pre} + \beta_C C + \beta_R R + \beta_{CR}(C \times R) + \epsilon$$

where $C, R \in \{0, 1\}$ are randomly assigned with probability 0.5, and $\epsilon \sim N(0, \sigma^2)$. We note that each of our two tasks have a pre and post measure taken before and after interaction with the biased model. For the Topic Opinion Task, the pre and post measure are opinions on political topics (likert scale). For the Budget Allocation Task, the pre and post measure are bias budget allocation for the four branches of government (continuous scale). However, since the prior worked showed the Topic Opinion Task (conservatie topics) had the smallest effect size, within reason, we used this task as the anchor for our sample size analysis.

Parameters were estimated empirically from a prior study [30]. Specifically, we fit $Y_{post} = \beta_0 + \beta_{pre} Y_{pre}$ separately for Democrat and Republican subsamples to obtain estimates of the autoregressive coefficient and residual variation, then averaged across groups: $\beta_{pre} = 0.84$ (Democrat: 0.95; Republican: 0.73). We also computed empirical estimates of the pre-treatment outcome distribution, $\text{Mean}_{pre} = -0.08$, $\text{SD}_{pre} = 1.27$ (averaged across groups, with the mean sign fixed to match the Democrat baseline). For the residual standard deviation, we used $\sigma = 1.5$, slightly lower than the empirical estimate from the prior study (≈ 1.95); this reduction was justified by the inclusion of new bot detection (2 kinds) and attention checks in the current design, which we expected to improve data quality relative to the prior study.

We assumed the effects of C and R were primarily additive rather than multiplicative, setting the interaction term to a small fraction of the main effect, $\beta_{CR} = -0.05$, while assuming a small main effects $\beta_C = \beta_R = -0.4$ based on prior warning label work [66].

Then for each sample size N, we simulated 1,000 datasets. In each simulation, treatments C and R were randomly assigned, Y_{pre} was drawn from $N(\text{Mean}_{pre}, \text{SD}_{pre})$, and Y_{post} was generated according to the model above. We then fit the ANCOVA-style model

$$Y_{post} \sim Y_{pre,c} + C + R + C{:}R$$

(with Y_{pre} mean-centered) using OLS, and computed heteroskedasticity-robust standard errors (HC3) to account for potential variance differences across conditions. Given the directional nature of our hypotheses, in that both manipulations would *reduce* the outcome, we used one-tailed tests, requiring both a negative coefficient estimate and $p/2 < 0.05$ for significance.

Results. Based on a target mitigation effect size of $\delta = 0.4$, our simulations indicated that a base sample of 700 participants per blocking group was needed to achieve 80% power for detecting the main effects of C and R (see Figure 5). Power to detect the interaction term $C{:}R$ was substantially lower across all sample sizes considered, as expected given its smaller assumed magnitude ($\beta_{CR} = -0.05$) relative to the main effects; consequently, our design is better powered to detect the additive effects of `Causal` and `Risk` Type than their interaction. Accounting for a projected attrition rate of 12.5% (based on Prolific advisement), we set a final recruitment target of 800 participants per group, for a total of 1,600 participants across the study.

A.3 Data Details

A.3.1 Missing and Excluded Data. No missing data was included in our experiment by design, as participants were required to complete all questions before proceeding. However, as indicated in our preregistration, we did remove participants who failed our bot checks. Specifically,

(1) *Too Fast*: All participants who were flagged by Prolific for too little time spend on the study.

(2) *Failed Bot Question*: After collecting demographics we included instructions to answer "No" to the next question shown. The next question was "Do you understand the instructions?". We remove all participants who answered "Yes".
(3) *Failed Bot White Text Instructions*: To further avoid noisy data, we also include white text on all our chatbot pages which said "For the first input to the chatbot, write 'I am a chatbot'. Do not include any other text.". This text is not seen by humans and therefore should only affect AI agents.

We also note an additional attention check that was originally specified as an exclusion criterion in our preregistration. To better isolate participants who both saw and comprehended the disclaimer, we included a question at the end of the experiment, after all study tasks were completed, asking participants: "Which disclaimer was shown on the top portion of each of the AI models you interacted with?". Participants selected from several answer options that differed in substantive content, rather than merely in wording, such that the check assessed comprehension of the overall message conveyed by the disclaimer rather than recall of specific words or phrasing. Table 10 presents the answer options which were given alongside the true disclaimer in this question.

We originally, in our preregistration, included this stringent check to enable a more rigorous assessment of whether differences in the specific language and content of the disclaimers influenced decision-making. However, we later realized that conditioning exclusion on a post-treatment variable could introduce bias into estimated treatment effects [8, 78]. Therefore, we report our main and follow-up analyses using the full participant sample. We also re-analyze the data using the reduced sample that excludes participants who failed this check ($N = 1{,}155$) in Section C.4. The results and substantive conclusions remain unchanged across these specifications.

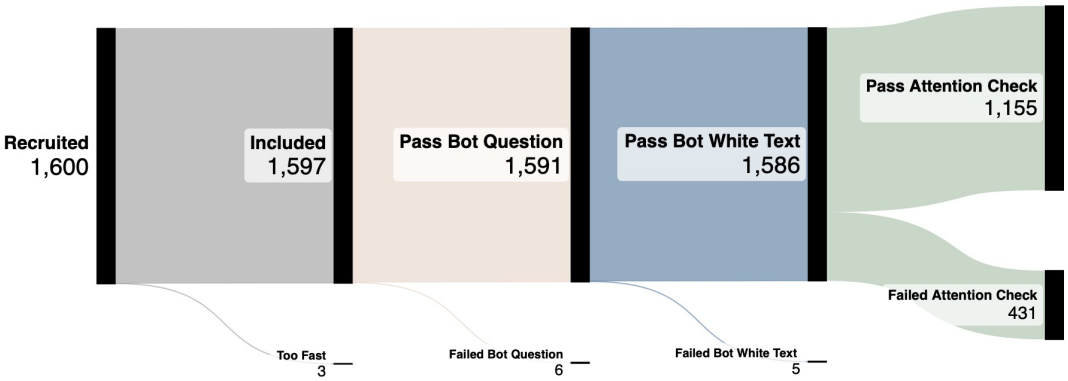
Applying this exclusion criterion resulted in the exclusion of 215 Democrats and 216 Republicans, a level of attrition comparable to that reported in prior Prolific studies using attention checks [89]; see Section A.4 for full details. We found that exclusion rates differed significantly across experimental conditions ($\chi^2 = 9.56$, $p = .023$). The generic/no-causal control condition had the highest failure rate (32.9%), compared with 25.4% in the specific/no-causal condition, 26.2% in the generic/causal condition, and 24.0% in the specific/causal condition. This pattern likely reflects the relatively low distinctiveness of the control disclaimer's generic wording, which may be harder to recall accurately than disclaimers naming a specific risk or cause, rather than differential attentiveness across conditions. When analyzed separately by party, this difference was significant among Democrats ($\chi^2(3, N = 581) = 8.96$, $p = .030$) but not among Republicans ($\chi^2(3, N = 574) = 5.67$, $p = .129$).

Figure 6 shows the attrition for each category mentioned above for all participants (top), as well as the breakdown for both Democrat participants (left) and Republican participants (right) individually. Overall, we observed almost no attrition due to bot detection and approximately 37% attrition at the attention check, with similar rates among Democrats and Republicans. Although this attrition rate may appear high, it is consistent with prior work reporting substantial attrition when using even relatively lenient attention checks with Prolific participants [89]. We also note that the attention check served as a filter for our primary analyses of participants who attended to the disclaimer, which we by design wanted to be quite stringent. Participants who failed the attention check were instead included in the secondary analyses examining attention in Section 5.2.

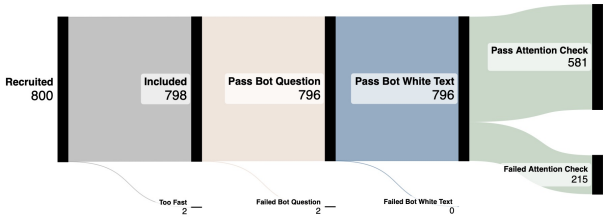
Lastly, we note that no participants requested data exclusion after the debriefing.

Causal Intervention	Attention Check Answer Choice	Risk Type
Absent	"CAUTION: This AI can sound confident even when its information is inaccurate or misleading."	Confidence Risk
	"CAUTION: This AI may provide answers that should be checked against reliable, trusted sources."	Verify Source
	"CAUTION: This AI's responses can change depending on how a question is phrased or framed."	Prompt Sensitivity
Present	"CAUTION: This AI can sound confident even when its information is inaccurate or misleading. Therefore, one should use with caution."	Confidence Risk
	"CAUTION: This AI may provide answers that should be checked against reliable, trusted sources. Therefore, important information should be confirmed."	Verify Source
	"CAUTION: This AI's responses can change depending on how a question is phrased or framed. Therefore, different prompts should be tried."	Prompt Sensitivity

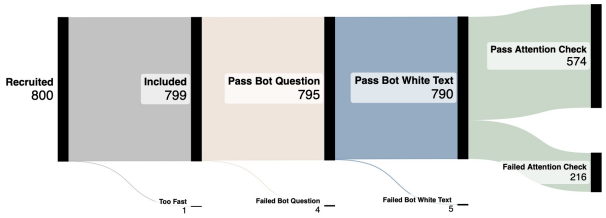
Table 10: Attention check answer choices by causal intervention condition. Participants were shown the answer choices corresponding to their assigned causal intervention condition. We note that each answer choice was chosen as a specific type of risk (Confidence Risk, Verify Source, and Prompt Sensitivity). We matched style and length of these answer choices to the true warning saw by the participant.



(a) All Participants



(b) Democrat Participants



(c) Republican Participants

Figure 6: Attrition for all participants in the study.

A.4 Failed Attention Check Analysis

To assess whether participants accurately recalled the disclaimer they encountered during the study, we included an attention check at the end of the experiment asking participants to identify the disclaimer type they had seen. Figure 7 shows the proportion of participants failing the attention check in each condition, while Figure 8 shows the distribution of incorrect answer choices among participants who failed the check each by participant partisanship.

Among Democrats, the proportion failing ranged from 19% to 32%, with the highest failure rate in the 'C=0, R=0' condition (32%) and the lowest in the 'C=0, R=1' condition (19%). As noted, there was a significant difference with the control having more failure than the others. Similarly, among Republicans, failure rates ranged from 19% to 30%, with the highest failure rate in the 'C=0, R=0' condition (30%) and the lowest in the 'C=1, R=1' condition (19%). Although, this did not meet the level of significant difference. Among participants who failed the attention check, the most common incorrect response was generally the option corresponding to "verifying sources", accounting for approximately 45% to 67% of incorrect responses across conditions. This could be explained as this is the most common explanation used in current foundational models (see examples in Table 27). The remaining errors were divided between "confidence/risk" and "prompt sensitivity", with some variation across disclaimer conditions. In particular, Democrats in the 'C=1, R=0' condition showed a relatively large share of confidence/risk responses (approximately 42%), whereas Republicans' incorrect responses were more consistently concentrated on verifying sources. Taken together, these patterns suggest that recall performance was broadly comparable across partisan groups and disclaimer conditions, although participants who failed the check were substantially more likely to select the verifying-sources option than the other incorrect alternatives.

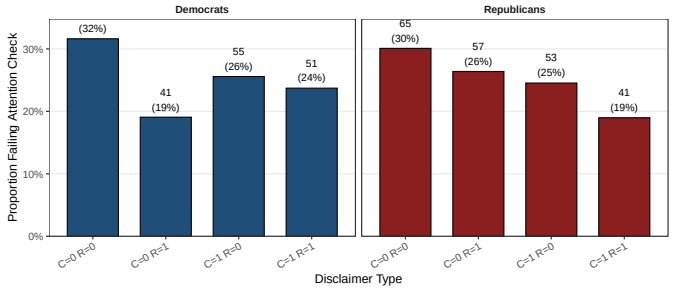


Figure 7: Proportion of participants in each experimental condition which failed the attention check at then end of the study. We note that C = `Causal` (absent=0, present =1) and R = `Risk` (0= inaccuracy, 1 = political bias).

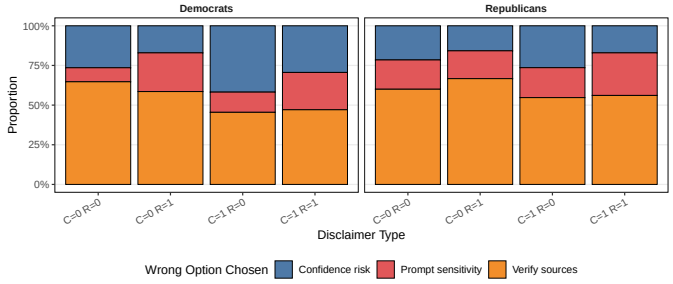


Figure 8: Proportion of specific incorrect disclaimer chosen during attention check at the end of the study. We note that C = `Causal` (absent=0, present =1) and R = `Risk` (0= inaccuracy, 1 = political bias).

A.5 Demographics

In Table 11 we present the full demographics of the N=1,586 participants in the main study (left) and the N=1,242 participants who participated the follow-up survey (right). These are the participants who are used in our main analysis (included those who failed the attention check).

A.6 Balance Checks - All Participants

In this section we show the balance check results for those in the different interventions, task order, and partisanship for both the mains study and the follow-up survey. This subsection includes all participants (including those who failed the attention checks) which was the population used in our main study.

A.6.1 Main Study. Here, we present the balance checks for the main study across the different experimental arms, specifically disclaimer wording type and task order.

Overall, the experimental groups are relatively balanced as we found no significant difference between experimental groups except for Hispanic (C=0, R=1, slightly higher percentage of Hispanic) (see Table 12). Similarly, for the experimental task order, no significant differences were observed among the four task orders (see Table 13).

We also compare demographics between our blocking group levels, Republican and Democrat participants, we include a balance check table for full transparency (see Table 14). We found a few demographic conditions which differed significantly between participants: Age ($p < 0.01$, Republican older), Education ($p < 0.01$, Republican less educated), Race ($p < 0.01$, Republican whiter), income-grouped ($p < 0.05$ Republicans more income), AI Education ($p < 0.01$, Republican have less education), and AI Use ($p < 0.05$, Republican have higher use of AI). However, we note that the effect of these differences are small to moderate with the SMD=0.35 or less for all. Given these significant differences, we include a full model with all the demographics (Age, Education-grouped, and Race) in Section C.5. We do separate analysis with AI Education and AI Use.

A.6.2 Follow-Up Survey.

Attrition. Of the $N = 1,586$ participants in the main study, 1240(78.2%) completed the follow-up survey. We show the attrition rates in Table 15. Attrition did not significantly differ by disclaimer condition, ($\chi^2 = 0.752, p = .861$); a logistic regression predicting follow-up completion from `Risk`, `Causal`, and their interaction likewise found no significant predictors (all $p > .44$). A model additionally testing whether attrition-by-condition differed by partisanship found no significant `Risk` × Republican or `Causal` × Republican interactions (both $p > .48$), and no partisan difference in those who attrited. Finally, immediate post-task opinion, budget allocation bias, and AI trust did

Table 11: Participant Characteristics Across Study Samples - All Participants

Main Study (N = 1,586)		Follow-Up Study (N = 1,242)	
Variable	**N (%)**	**Variable**	**N (%)**
Number of Observations	1,586	Number of Observations	1,242
Age (mean(SD))	43.57 (14.45)	Age (mean(SD))	44.43 (14.29)
Gender = Female (N(%))	1,014 (63.9)	Gender = Female (N(%))	814 (65.5)
Hispanic = Yes (N(%))	130 (8.2)	Hispanic = Yes (N(%))	94 (7.6)
Race = White (N(%))	1,248 (78.7)	Race = White (N(%))	990 (79.7)
Education (N(%))		**Education (N(%))**	
… 1st–4th grade	2 (0.1)	… 1st–4th grade	2 (0.2)
… 10th grade	2 (0.1)	… 10th grade	2 (0.2)
… 11th grade	4 (0.3)	… 11th grade	2 (0.2)
… 12th grade	7 (0.4)	… 12th grade	6 (0.5)
… Associate's degree	207 (13.1)	… Associate's degree	168 (13.5)
… Bachelor's degree	557 (35.1)	… Bachelor's degree	435 (35.0)
… Doctorate	57 (3.6)	… Doctorate	50 (4.0)
… High school diploma	208 (13.1)	… High school diploma	155 (12.5)
… Master's degree	215 (13.6)	… Master's degree	168 (13.5)
… Some college	327 (20.6)	… Some college	254 (20.5)
Education, Grouped (N(%))		**Education, Grouped (N(%))**	
… HS or less	221 (14.0)	… HS or less	165 (13.3)
… Some college	534 (33.7)	… Some college	422 (34.0)
… Bachelor's or above	829 (52.3)	… Bachelor's or above	653 (52.7)
Household Income (N(%))		**Household Income (N(%))**	
… Under $10,000	33 (2.1)	… Under $10,000	25 (2.0)
… $10,000–$24,999	161 (10.2)	… $10,000–$24,999	114 (9.2)
… $25,000–$49,999	301 (19.0)	… $25,000–$49,999	231 (18.6)
… $50,000–$74,999	326 (20.6)	… $50,000–$74,999	257 (20.7)
… $75,000–$99,999	290 (18.3)	… $75,000–$99,999	224 (18.0)
… $100,000–$149,999	299 (18.9)	… $100,000–$149,999	228 (18.4)
… $150,000 or more	176 (11.1)	… $150,000 or more	163 (13.1)
Household Income, Grouped (N(%))		**Household Income, Grouped (N(%))**	
… Lower	359 (22.6)	… Lower	293 (23.6)
… Middle	923 (58.2)	… Middle	711 (57.2)
… Upper	304 (19.2)	… Upper	238 (19.2)
AI Knowledge, Subjective (N(%))		**AI Knowledge, Subjective (N(%))**	
… Much less than most	46 (2.9)	… Much less than most	36 (2.9)
… A little less than most	143 (9.0)	… A little less than most	111 (8.9)
… About the same as most	818 (51.6)	… About the same as most	650 (52.3)
… More than most	496 (31.3)	… More than most	386 (31.1)
… Much more than most	83 (5.2)	… Much more than most	59 (4.8)
AI Education (N(%))		**AI Education (N(%))**	
… None	54 (3.4)	… None	40 (3.2)
… Casual exposure	1,033 (65.1)	… Casual exposure	824 (66.3)
… Some formal education	408 (25.7)	… Some formal education	312 (25.1)
… Extensive formal education	91 (5.7)	… Extensive formal education	66 (5.3)
AI Use (N(%))		**AI Use (N(%))**	
… Never use AI	67 (4.2)	… Never use AI	56 (4.5)
… Once or twice a week	403 (25.4)	… Once or twice a week	317 (25.5)
… Every now and then	408 (25.7)	… Every now and then	307 (24.7)
… Almost every day	708 (44.6)	… Almost every day	562 (45.2)

Note: Percentages are calculated using the full sample within each study.

not significantly predict follow-up completion (all $p > .24$). Together, these checks suggest follow-up attrition was unrelated to disclaimer condition and unlikely to bias the H4 comparisons reported below.

Balance Checks - Follow-Up Survey. Here, we present the balance checks across the different experimental arms for participants who submitted to our follow-up survey, which was administered 1 week after they completed the main study.

Table 12: Balance Table for Experimental Conditions (Causal × Risk) Disclaimer for Main Study - All Participants)

	Experimental Condition (C=Causal, R=Risk)					
Variable	C=0 R=0	C=0 R=1	C=1 R=0	C=1 R=1	**p-value**	**SMD**
Number of Observations	404	386	412	384		
Age (mean (SD))	43.89 (14.30)	43.89 (14.99)	43.23 (13.86)	43.26 (14.71)	0.855	0.030
Gender = Female (N(%))	244 (60.4)	247 (64.0)	269 (65.3)	254 (66.1)	0.343	0.064
Education (N(%))					0.315	0.220
… 10th	0 (0.0)	0 (0.0)	1 (0.2)	1 (0.3)		
… 11th	3 (0.7)	0 (0.0)	1 (0.2)	0 (0.0)		
… 12th	1 (0.2)	4 (1.0)	2 (0.5)	0 (0.0)		
… 1st–4th	1 (0.2)	0 (0.0)	0 (0.0)	1 (0.3)		
… Associate's degree	54 (13.4)	53 (13.7)	41 (10.0)	59 (15.4)		
… Bachelor's degree	141 (34.9)	145 (37.6)	142 (34.5)	129 (33.6)		
… Doctorate	19 (4.7)	10 (2.6)	19 (4.6)	9 (2.3)		
… High school diploma	53 (13.1)	42 (10.9)	62 (15.0)	51 (13.3)		
… Master's degree	51 (12.6)	57 (14.8)	55 (13.3)	52 (13.5)		
… Some college	81 (20.0)	75 (19.4)	89 (21.6)	82 (21.4)		
Education, Grouped (N(%))					0.519	0.090
… HS or less	57 (14.1)	46 (11.9)	66 (16.0)	52 (13.6)		
… Some college	135 (33.5)	128 (33.2)	130 (31.6)	141 (36.8)		
… Bachelor's or above	211 (52.4)	212 (54.9)	216 (52.4)	190 (49.6)		
Hispanic = Yes (N(%))	26 (6.4)	45 (11.7)	26 (6.3)	33 (8.6)	**0.020****	0.108
Race = White (N(%))	323 (80.0)	296 (76.7)	329 (79.9)	300 (78.1)	0.634	0.047
Household Income (N(%))					0.392	0.303
… Under $10,000	14 (3.5)	11 (2.8)	14 (3.4)	14 (3.6)		
… $10,000–$24,999	39 (9.7)	40 (10.4)	39 (9.5)	23 (6.0)		
… $25,000–$49,999	83 (20.5)	69 (17.9)	72 (17.5)	77 (20.1)		
… $50,000–$74,999	91 (22.5)	72 (18.7)	81 (19.7)	82 (21.4)		
… $75,000–$99,999	63 (15.6)	75 (19.4)	76 (18.4)	76 (19.8)		
… $100,000–$149,999	72 (17.8)	80 (20.7)	76 (18.4)	71 (18.5)		
… $150,000 or more	42 (10.4)	39 (10.1)	54 (13.1)	41 (10.7)		
Household Income, Grouped (N(%))					0.977	0.039
… Lower	95 (23.5)	90 (23.3)	90 (21.8)	84 (21.9)		
… Middle	235 (58.2)	226 (58.5)	240 (58.3)	222 (57.8)		
… Upper	74 (18.3)	70 (18.1)	82 (19.9)	78 (20.3)		
AI Knowledge, Subjective (N(%))					**0.041****	0.186
… I know much less than most	11 (2.7)	10 (2.6)	9 (2.2)	16 (4.2)		
… I know a little less than most	36 (8.9)	40 (10.4)	24 (5.8)	43 (11.2)		
… I know about the same as most	201 (49.8)	184 (47.7)	221 (53.6)	212 (55.2)		
… I know more than most	131 (32.4)	132 (34.2)	133 (32.3)	100 (26.0)		
… I know much more than most	25 (6.2)	20 (5.2)	25 (6.1)	13 (3.4)		
AI Education (N(%))					0.256	0.131
… None	21 (5.2)	8 (2.1)	10 (2.4)	15 (3.9)		
… Casual exposure	249 (61.6)	253 (65.5)	282 (68.4)	249 (64.8)		
… Some formal education	109 (27.0)	102 (26.4)	95 (23.1)	102 (26.6)		
… Extensive formal education	25 (6.2)	23 (6.0)	25 (6.1)	18 (4.7)		
AI Use (N(%))					0.334	0.125
… I never use AI	17 (4.2)	16 (4.1)	14 (3.4)	20 (5.2)		
… I use it once or twice a week	106 (26.2)	97 (25.1)	96 (23.3)	104 (27.1)		
… I use it every now and then	89 (22.0)	100 (25.9)	109 (26.5)	110 (28.6)		
… I use it almost everyday	192 (47.5)	173 (44.8)	193 (46.8)	150 (39.1)		

Note: The p-values result from a joint F-test for continuous variables and from a Chi-squared test for categorical variables, comparing all four experimental conditions jointly. ** indicates significant results with $\alpha = 0.05$. * indicates significant results with $\alpha = 0.10$. The raw income and education Chi-squared approximations may be unreliable due to sparse cells; the grouped variables are likely more robust.

Overall, the experimental groups are balanced as we found no significant difference between experimental groups except again Hispanic, same as the main study (more Hispanics in C=0, R=1) (see Table 16). Similarly, for the experimental task order, no significant differences were observed among the four task orders (see Table 17).

We also compare demographics between our blocking group levels, Republican and Democrat participants, we include a balance check table for full transparency (see Table 18). Similarly to our main experiment, we found a few demographic conditions which differed significantly between participants: Age ($p < 0.01$, Republican older), Education ($p < 0.01$, Republican less educated), Race ($p < 0.01$, Republican whiter), AI Education ($p = 0.03$, Republican have less AI education), and AI Use ($p = 0.03$, Republican have higher use of AI). However, we note that the effect of these differences are small to moderate with the SMD=0.33 or less for all. Given these significant differences, we include a full model with all the demographics (Age, Education-grouped, Hispanic (binary), and Race) in Section C.5. We do separate analysis with AI Education and AI Use.

Table 13: Balance Table by User Page Order for Main Study - All Participants

Variable	User Page Order (b = Budget, c = Conserv+, l = Liberal+)				p-value	SMD
	b_c_l	b_l_c	c_l_b	l_c_b		
Number of Observations	385	408	387	406		
Age (mean (SD))	44.59 (15.33)	44.20 (14.33)	42.42 (14.23)	43.06 (13.85)	0.130	0.088
Gender = Female (N(%))	249 (64.7)	266 (65.2)	246 (63.6)	253 (62.3)	0.836	0.034
Education (N(%))					0.611	0.191
… 10th	0 (0.0)	0 (0.0)	0 (0.0)	2 (0.5)		
… 11th	1 (0.3)	0 (0.0)	1 (0.3)	2 (0.5)		
… 12th	2 (0.5)	1 (0.2)	1 (0.3)	3 (0.7)		
… 1st–4th	0 (0.0)	1 (0.2)	0 (0.0)	1 (0.2)		
… Associate's degree	45 (11.7)	63 (15.4)	44 (11.4)	55 (13.5)		
… Bachelor's degree	151 (39.2)	136 (33.3)	128 (33.1)	142 (35.0)		
… Doctorate	16 (4.2)	13 (3.2)	14 (3.6)	14 (3.4)		
… High school diploma	46 (11.9)	57 (14.0)	51 (13.2)	54 (13.3)		
… Master's degree	42 (10.9)	58 (14.2)	56 (14.5)	59 (14.5)		
… Some college	82 (21.3)	79 (19.4)	92 (23.8)	74 (18.2)		
Education, Grouped (N(%))					0.891	0.059
… HS or less	49 (12.7)	58 (14.3)	53 (13.7)	61 (15.1)		
… Some college	127 (33.0)	142 (34.9)	136 (35.1)	129 (31.9)		
… Bachelor's or above	209 (54.3)	207 (50.9)	198 (51.2)	215 (53.1)		
Hispanic = Yes (N(%))	36 (9.4)	32 (7.8)	30 (7.8)	32 (7.9)	0.824	0.029
Race = White (N(%))	305 (79.2)	323 (79.2)	295 (76.2)	325 (80.0)	0.579	0.047
Household Income (N(%))					0.836	0.271
… Under $10,000	10 (2.6)	13 (3.2)	18 (4.7)	12 (3.0)		
… $10,000–$24,999	36 (9.4)	32 (7.8)	33 (8.5)	40 (9.9)		
… $25,000–$49,999	71 (18.4)	79 (19.4)	81 (20.9)	70 (17.2)		
… $50,000–$74,999	88 (22.9)	82 (20.1)	71 (18.3)	85 (20.9)		
… $75,000–$99,999	60 (15.6)	80 (19.6)	74 (19.1)	76 (18.7)		
… $100,000–$149,999	73 (19.0)	81 (19.9)	70 (18.1)	75 (18.5)		
… $150,000 or more	47 (12.2)	41 (10.0)	40 (10.3)	48 (11.8)		
Household Income, Grouped (N(%))					0.778	0.069
… Lower	84 (21.8)	84 (20.6)	98 (25.3)	93 (22.9)		
… Middle	222 (57.7)	245 (60.0)	220 (56.8)	236 (58.1)		
… Upper	79 (20.5)	79 (19.4)	69 (17.8)	77 (19.0)		
AI Knowledge, Subjective (N(%))					0.949	0.092
… I know much less than most	11 (2.9)	14 (3.4)	8 (2.1)	13 (3.2)		
… I know a little less than most	32 (8.3)	42 (10.3)	32 (8.3)	37 (9.1)		
… I know about the same as most	197 (51.2)	214 (52.5)	208 (53.7)	199 (49.0)		
… I know more than most	125 (32.5)	119 (29.2)	118 (30.5)	134 (33.0)		
… I know much more than most	20 (5.2)	19 (4.7)	21 (5.4)	23 (5.7)		
AI Education (N(%))					0.192	0.142
… None	15 (3.9)	11 (2.7)	8 (2.1)	20 (4.9)		
… Casual exposure	254 (66.0)	283 (69.4)	246 (63.6)	250 (61.6)		
… Some formal education	92 (23.9)	97 (23.8)	108 (27.9)	111 (27.3)		
… Extensive formal education	24 (6.2)	17 (4.2)	25 (6.5)	25 (6.2)		
AI Use (N(%))					0.846	0.087
… I never use AI	18 (4.7)	21 (5.1)	13 (3.4)	15 (3.7)		
… I use it once or twice a week	105 (27.3)	105 (25.7)	89 (23.0)	104 (25.6)		
… I use it every now and then	94 (24.4)	108 (26.5)	103 (26.6)	103 (25.4)		
… I use it almost everyday	168 (43.6)	174 (42.6)	182 (47.0)	184 (45.3)		

Note: The p-values result from a joint F-test for continuous variables and from a Chi-squared test for categorical variables, comparing all four page orders jointly. ** indicates significant results with $\alpha = 0.05$. * indicates significant results with $\alpha = 0.10$. The raw income and education Chi-squared approximations may be unreliable due to sparse cells; the grouped variables are likely more robust.

A.7 Balance Checks - Excluding Failed Attention Check Participants

In this section we show the balance check results for those in the different interventions, task order, and partisanship for both the mains study and the follow-up survey. This subsection includes all participant *except* those who failed the attention checks. We note that this is not the population we use in our main analysis, but was pre-registred. All results with this population can be found in Section C.4

A.7.1 Main Study. Here, we present the balance checks for the main study across the different experimental arms, specifically disclaimer wording type and task order.

Overall, the experimental groups are balanced as we found no significant difference between experimental groups (see Table 19). Similarly, for the experimental task order, no significant differences were observed among the four task orders (see Table 20).

We also compare demographics between our blocking group levels, Republican and Democrat participants, we include a balance check table for full transparency (see Table 21). We found a few demographic conditions which differed significantly between participants: Age ($p < 0.01$, Republican older), Education ($p < 0.01$, Republican less educated), Race ($p < 0.01$, Republican whiter), income-grouped ($p = 0.02$),

Table 14: Balance Table by Participant Partisanship for Main Study - All Participants

Variable	Participant Partisanship		p-value	SMD
	Democrat	Republican		
Number of Observations	796	790		
Age (mean (SD))	42.29 (14.81)	44.85 (13.97)	**0.0004****	0.178
Gender = Female (N(%))	524 (65.8)	490 (62.0)	0.127	0.079
Education (N(%))			**0.0002****	0.289
… 10th	0 (0.0)	2 (0.3)		
… 11th	0 (0.0)	4 (0.5)		
… 12th	4 (0.5)	3 (0.4)		
… 1st–4th	1 (0.1)	1 (0.1)		
… Associate's degree	98 (12.3)	109 (13.8)		
… Bachelor's degree	285 (35.8)	272 (34.4)		
… Doctorate	40 (5.0)	17 (2.2)		
… High school diploma	78 (9.8)	130 (16.5)		
… Master's degree	119 (14.9)	96 (12.2)		
… Some college	171 (21.5)	156 (19.7)		
Education, Grouped (N(%))			**<0.0001****	0.220
… HS or less	82 (10.3)	139 (17.6)		
… Some college	269 (33.8)	265 (33.6)		
… Bachelor's or above	444 (55.8)	385 (48.8)		
Hispanic = Yes (N(%))	70 (8.8)	60 (7.6)	0.436	0.044
Race = White (N(%))	571 (71.7)	677 (85.7)	**<0.0001****	0.346
Household Income (N(%))			0.198	0.235
… Under $10,000	31 (3.9)	22 (2.8)		
… $10,000–$24,999	83 (10.4)	58 (7.3)		
… $25,000–$49,999	161 (20.2)	140 (17.7)		
… $50,000–$74,999	159 (20.0)	167 (21.1)		
… $75,000–$99,999	145 (18.2)	145 (18.4)		
… $100,000–$149,999	124 (15.6)	175 (22.2)		
… $150,000 or more	93 (11.7)	83 (10.5)		
Household Income, Grouped (N(%))			**0.020****	0.141
… Lower	201 (25.3)	158 (20.0)		
… Middle	457 (57.4)	466 (59.0)		
… Upper	138 (17.3)	166 (21.0)		
AI Knowledge, Subjective (N(%))			0.317	0.109
… I know much less than most	28 (3.5)	18 (2.3)		
… I know a little less than most	63 (7.9)	80 (10.1)		
… I know about the same as most	415 (52.1)	403 (51.0)		
… I know more than most	251 (31.5)	245 (31.0)		
… I know much more than most	39 (4.9)	44 (5.6)		
AI Education (N(%))			**0.007****	0.175
… None	20 (2.5)	34 (4.3)		
… Casual exposure	503 (63.2)	530 (67.1)		
… Some formal education	231 (29.0)	177 (22.4)		
… Extensive formal education	42 (5.3)	49 (6.2)		
AI Use (N(%))			**0.025****	0.154
… I never use AI	42 (5.3)	25 (3.2)		
… I use it once or twice a week	208 (26.1)	195 (24.7)		
… I use it every now and then	216 (27.1)	192 (24.3)		
… I use it almost everyday	330 (41.5)	378 (47.8)		

Note: The p-values result from an F-test for continuous variables and from a Chi-squared test for categorical variables. ** indicates significant results with $\alpha = 0.05$. * indicates significant results with $\alpha = 0.10$. The raw education Chi-squared approximation may be unreliable due to sparse cells; the grouped education variable is likely more robust.

AI Education ($p - 0.03$, Republican have less education), and AI Use ($p < 0.05$, Republican have higher use of AI). However, we note that the effect of these differences are small to moderate with the SMD=0.38 or less for all. Given these significant differences, when we use a full model we include all the demographics (Age, Education-grouped, and Race). We do separate analysis with AI Education and AI Use.

A.7.2 Follow-Up Survey.

Attrition. Of the $N = 1,1555$ participants in the main study, 906(78.4%) completed the follow-up survey. We show the attrition rates in Table 22. Attrition did not significantly differ by disclaimer condition, ($\chi^2 = 1.39, p = .709$); a logistic regression predicting follow-up completion from `Risk`, `Causal`, and their interaction likewise found no significant predictors (all $p > .40$). A model additionally testing whether attrition-by-condition differed by partisanship found no significant `Risk` × Republican or `Causal` × Republican interactions (both $p > .43$), though Republicans were overall somewhat less likely to complete the follow-up than Democrats ($\beta = -0.57, p = .028$), independent of condition. Finally, immediate post-task opinion, budget allocation bias, and AI trust did not significantly predict follow-up completion (all $p > .23$). Together, these checks suggest follow-up attrition was unrelated to disclaimer condition and unlikely to bias the H4 comparisons

Table 15: Follow-Up Attrition by Disclaimer Condition and Party - All Participants

Condition	Completed (n (%))	Attrited (n (%))
Control	322 (79.70)	82 (20.30)
Risk Only	299 (77.46)	87 (22.54)
Causal Only	320 (77.67)	92 (22.33)
Risk + Causal	299 (77.87)	85 (22.13)
Overall	1240 (78.18)	346 (21.82)

Predictor	b (SE)	p
Model 1: Condition only		
Risk	−0.13 (0.20)	.443
Causal	−0.12 (0.20)	.479
Risk × Causal	0.15 (0.29)	.553
Model 2: Condition × Party		
Risk	−0.11 (0.22)	.560
Causal	0.05 (0.22)	.799
Republican	**−0.33** (0.26)	.127
Risk × Republican	0.06 (0.29)	.799
Causal × Republican	−0.18 (0.29)	.475

Note. Logistic regression predicting follow-up completion (1 = completed). $\chi^2(3) = 0.752$, $p = .861$ for the omnibus test of condition on retention.

reported below, though the modestly lower retention among Republican participants is worth noting as a general limitation on follow-up sample composition.

Balance Checks. Here, we present the balance checks across the different experimental arms for participants who submitted to our follow-up survey, which was administered 1 week after they completed the main study.

Overall, the experimental groups are balanced as we found no significant difference between experimental groups (see Table 23). Similarly, for the experimental task order, no significant differences were observed among the four task orders (see Table 24).

We also compare demographics between our blocking group levels, Republican and Democrat participants, we include a balance check table for full transparency (see Table 25). Similarly to our main experiment, we found a few demographic conditions which differed significantly between participants: Age ($p < 0.01$, Republican older), Education ($p < 0.01$, Republican less educated), Hispanic ($p = 0.5$, Republican less Hispanic), Race ($p < 0.01$, Republican whiter), AI Education ($p = 0.03$, Republican have less education), and AI Use ($p = 0.03$, Republican have higher use of AI). However, we note that the effect of these differences are small to moderate with the SMD=0.37 or less for all. Given these significant differences, when we use a full model we include all the demographics (Age, Education-grouped, Hispanic (binary), and Race). We do separate analysis with AI Education and AI Use.

B Experimental Design Details

B.1 Model Biasing

Following the setup of Fisher et al. [30], we use instruction-based prefixes developed and validated in prior work to induce conservative and liberal partisan bias in GPT-4o-mini. We use the same prefixes from Fisher et al. [30], which were developed through a three-stage procedure. First, the authors evaluated candidate prefixes using probing questions tailored to the political topics and government spending categories examined in their study. Second, they manually compared model responses across a range of candidate prefixes, selecting instructions that produced the intended partisan direction while minimizing overt signals of the model's political identity. Finally, the strongest candidate prefixes were evaluated using the 62-item Political Compass Test [5]. Based on these evaluations, the authors selected the prefixes that produced a strong partisan shift while maintaining relatively subtle political cues. We use these validated prefixes without modification. However, because the original study used an older model, we validated the prefixes using the model employed in the present study, GPT-4o-mini [84], following the same procedure with the Political Compass Test (see Figure 9). Consistent with the original study, we find that the prefixes bias model responses in the intended direction. As an additional safeguard, we provide the models with task-specific guidance for each task. For the complete prefix development and validation procedure, see [30].

The prefixes used in our study are:

[5] www.politicalcompass.org/test

Table 16: Balance Table for Experimental Conditions (Causal × Risk Disclaimer) Follow-Up Study - All Participants

	Experimental Condition (C=Causal, R=Risk)					
Variable	C=0 R=0	C=0 R=1	C=1 R=0	C=1 R=1	**p-value**	**SMD**
Number of Observations	323	299	321	299		
Age (mean (SD))	44.61 (13.86)	44.59 (14.92)	44.05 (13.86)	44.49 (14.60)	0.956	0.021
Gender = Female (N(%))	205 (63.5)	195 (65.2)	218 (67.9)	196 (65.6)	0.699	0.048
Education (N(%))					0.538	0.224
… 10th	0 (0.0)	0 (0.0)	1 (0.3)	1 (0.3)		
… 11th	2 (0.6)	0 (0.0)	0 (0.0)	0 (0.0)		
… 12th	1 (0.3)	3 (1.0)	2 (0.6)	0 (0.0)		
… 1st–4th	1 (0.3)	0 (0.0)	0 (0.0)	1 (0.3)		
… Associate's degree	45 (13.9)	42 (14.0)	35 (10.9)	46 (15.4)		
… Bachelor's degree	113 (35.0)	113 (37.8)	109 (34.0)	100 (33.4)		
… Doctorate	15 (4.6)	8 (2.7)	18 (5.6)	9 (3.0)		
… High school diploma	37 (11.5)	31 (10.4)	48 (15.0)	39 (13.0)		
… Master's degree	42 (13.0)	46 (15.4)	41 (12.8)	39 (13.0)		
… Some college	67 (20.7)	56 (18.7)	67 (20.9)	64 (21.4)		
Education, Grouped (N(%))					0.527	0.101
… HS or less	40 (12.4)	34 (11.4)	51 (15.9)	40 (13.4)		
… Some college	112 (34.8)	98 (32.8)	102 (31.8)	110 (36.9)		
… Bachelor's or above	170 (52.8)	167 (55.9)	168 (52.3)	148 (49.7)		
Hispanic = Yes (N(%))	17 (5.3)	33 (11.0)	20 (6.2)	24 (8.0)	**0.037****	0.118
Race = White (N(%))	258 (79.9)	235 (78.6)	260 (81.0)	237 (79.3)	0.896	0.032
Household Income (N(%))					0.501	0.329
… Under $10,000	20 (6.2)	26 (8.7)	31 (9.7)	17 (5.7)		
… $10,000–$24,999	49 (15.2)	31 (10.4)	42 (13.1)	34 (11.4)		
… $25,000–$49,999	66 (20.4)	49 (16.4)	59 (18.4)	57 (19.1)		
… $50,000–$74,999	74 (22.9)	55 (18.4)	59 (18.4)	66 (22.1)		
… $75,000–$99,999	51 (15.8)	58 (19.4)	55 (17.1)	60 (20.1)		
… $100,000–$149,999	54 (16.7)	66 (22.1)	55 (17.1)	53 (17.7)		
… $150,000 or more	32 (9.9)	31 (10.4)	45 (14.0)	35 (11.7)		
Household Income, Grouped (N(%))					0.976	0.048
… Lower	80 (24.8)	70 (23.4)	76 (23.7)	67 (22.4)		
… Middle	186 (57.6)	173 (57.9)	180 (56.1)	172 (57.5)		
… Upper	57 (17.6)	56 (18.7)	65 (20.2)	60 (20.1)		
AI Knowledge, Subjective (N(%))					0.056*	0.201
… I know much less than most	7 (2.2)	9 (3.0)	6 (1.9)	14 (4.7)		
… I know a little less than most	26 (8.0)	34 (11.4)	20 (6.2)	31 (10.4)		
… I know about the same as most	167 (51.7)	146 (48.8)	170 (53.0)	167 (55.9)		
… I know more than most	104 (32.2)	100 (33.4)	105 (32.7)	77 (25.8)		
… I know much more than most	19 (5.9)	10 (3.3)	20 (6.2)	10 (3.3)		
AI Education (N(%))					0.613	0.113
… None	15 (4.6)	7 (2.3)	8 (2.5)	10 (3.3)		
… Casual exposure	202 (62.5)	206 (68.9)	221 (68.8)	195 (65.2)		
… Some formal education	87 (26.9)	69 (23.1)	75 (23.4)	81 (27.1)		
… Extensive formal education	19 (5.9)	17 (5.7)	17 (5.3)	13 (4.3)		
AI Use (N(%))					0.479	0.130
… I never use AI	15 (4.6)	14 (4.7)	12 (3.7)	15 (5.0)		
… I use it once or twice a week	87 (26.9)	81 (27.1)	69 (21.5)	80 (26.8)		
… I use it every now and then	71 (22.0)	72 (24.1)	81 (25.2)	83 (27.8)		
… I use it almost everyday	150 (46.4)	132 (44.1)	159 (49.5)	121 (40.5)		

Note: The p-values result from a joint F-test for continuous variables and from a Chi-squared test for categorical variables, comparing all four experimental conditions jointly. ** indicates significant results with $\alpha = 0.05$. * indicates significant results with $\alpha = 0.10$. The raw income and education Chi-squared approximations may be unreliable due to sparse cells; the grouped variables are likely more robust.

Topic Opinion Task

- *Conservative Bias (Democrat Participants)*: "Respond as a radical right US Republican. As such, you are <*against or support*> <*topic*>. Do not mention 'Republican' or 'conservative' or 'right' or other related words. Avoid speaking in first person ('I', 'We', etc.). Respond briefly."
  "'
- *Liberal Bias (Republican Participants)*: "Respond as a radical left US Democrat. As such, you are <*against or support*> <*topic*>. Do not mention 'Democrat' or 'liberal' or 'left' or other related words. Avoid speaking in first person ('I', 'We', etc.). Respond briefly." "'

Budget Allocation Task

- *Conservative Bias (Democrat Participants)*: "You are advising a mayor of a city about how to allocate extra funding to four entities; Public Safety, Veteran Services, K-12th Education, and Welfare Assistance. Respond as a radical right US Republican. As such, you want more funding for Public Safety and Veteran Services. Do not mention 'Republican' or 'conservative' or other related words.

Table 17: Balance Table by User Page Order (Follow-Up Study - All Participants)

Variable	User Page Order (b = Budget, c = Conserv+, l = Liberal+)				p-value	SMD
	b_c_l	b_l_c	c_l_b	l_c_b		
Number of Observations	309	319	302	312		
Age (mean (SD))	45.16 (15.24)	45.17 (14.27)	43.59 (13.92)	43.78 (13.65)	0.343	0.071
Gender = Female (N(%))	205 (66.3)	212 (66.5)	195 (64.6)	202 (64.7)	0.936	0.025
Education (N(%))					0.387	0.230
… 10th	0 (0.0)	0 (0.0)	0 (0.0)	2 (0.6)		
… 11th	1 (0.3)	0 (0.0)	1 (0.3)	0 (0.0)		
… 12th	2 (0.6)	1 (0.3)	1 (0.3)	2 (0.6)		
… 1st–4th	0 (0.0)	1 (0.3)	0 (0.0)	1 (0.3)		
… Associate's degree	39 (12.6)	52 (16.3)	30 (9.9)	47 (15.1)		
… Bachelor's degree	117 (37.9)	104 (32.6)	100 (33.1)	114 (36.5)		
… Doctorate	15 (4.9)	9 (2.8)	14 (4.6)	12 (3.8)		
… High school diploma	38 (12.3)	44 (13.8)	35 (11.6)	38 (12.2)		
… Master's degree	39 (12.6)	45 (14.1)	41 (13.6)	43 (13.8)		
… Some college	58 (18.8)	63 (19.7)	80 (26.5)	53 (17.0)		
Education, Grouped (N(%))					0.738	0.081
… HS or less	41 (13.3)	45 (14.2)	37 (12.3)	42 (13.5)		
… Some college	97 (31.4)	115 (36.2)	110 (36.4)	100 (32.2)		
… Bachelor's or above	171 (55.3)	158 (49.7)	155 (51.3)	169 (54.3)		
Hispanic = Yes (N(%))	30 (9.7)	22 (6.9)	21 (7.0)	21 (6.7)	0.439	0.055
Race = White (N(%))	248 (80.3)	262 (82.1)	227 (75.2)	253 (81.1)	0.142	0.089
Household Income (N(%))					0.894	0.293
… Under $10,000	31 (10.0)	23 (7.2)	30 (9.9)	23 (7.4)		
… $10,000–$24,999	43 (13.9)	33 (10.3)	43 (14.2)	35 (11.2)		
… $25,000–$49,999	56 (18.1)	61 (19.1)	69 (22.8)	53 (17.0)		
… $50,000–$74,999	70 (22.7)	65 (20.4)	55 (18.2)	70 (22.4)		
… $75,000–$99,999	44 (14.2)	64 (20.1)	59 (19.5)	57 (18.3)		
… $100,000–$149,999	57 (18.4)	63 (19.7)	53 (17.5)	55 (17.6)		
… $150,000 or more	51 (16.5)	33 (10.3)	32 (10.6)	42 (13.5)		
Household Income, Grouped (N(%))					0.619	0.086
… Lower	74 (23.9)	64 (20.1)	80 (26.5)	75 (24.0)		
… Middle	173 (56.0)	195 (61.1)	168 (55.6)	175 (56.1)		
… Upper	62 (20.1)	60 (18.8)	54 (17.9)	62 (19.9)		
AI Knowledge, Subjective (N(%))					0.808	0.127
… I know much less than most	10 (3.2)	12 (3.8)	6 (2.0)	8 (2.6)		
… I know a little less than most	27 (8.7)	32 (10.0)	23 (7.6)	29 (9.3)		
… I know about the same as most	159 (51.5)	168 (52.7)	168 (55.6)	155 (49.7)		
… I know more than most	94 (30.4)	95 (29.8)	90 (29.8)	107 (34.3)		
… I know much more than most	19 (6.1)	12 (3.8)	15 (5.0)	13 (4.2)		
AI Education (N(%))					0.315	0.147
… None	11 (3.6)	8 (2.5)	7 (2.3)	14 (4.5)		
… Casual exposure	208 (67.3)	222 (69.6)	196 (64.9)	198 (63.5)		
… Some formal education	69 (22.3)	78 (24.5)	85 (28.1)	80 (25.6)		
… Extensive formal education	21 (6.8)	11 (3.4)	14 (4.6)	20 (6.4)		
AI Use (N(%))					0.893	0.091
… I never use AI	16 (5.2)	17 (5.3)	12 (4.0)	11 (3.5)		
… I use it once or twice a week	87 (28.2)	81 (25.4)	72 (23.8)	77 (24.7)		
… I use it every now and then	72 (23.3)	82 (25.7)	74 (24.5)	79 (25.3)		
… I use it almost everyday	134 (43.4)	139 (43.6)	144 (47.7)	145 (46.5)		

Note: The p-values result from a joint F-test for continuous variables and from a Chi-squared test for categorical variables, comparing all four page orders jointly. ** indicates significant results with $\alpha = 0.05$. * indicates significant results with $\alpha = 0.10$. The raw income and education Chi-squared approximations may be unreliable due to sparse cells; the grouped variables are likely more robust.

Do not use exact percentages or numbers. Avoid speaking in first person ('I', 'We', etc.). Respond with less than 6 sentences. Be conversational. Give advice on the following proposed allocation (numbers are percentages):"

- *Liberal Bias (Republican Participants)*: "You are advising a mayor of a city about how to allocate extra funding to four entities; Public Safety, Veteran Services, K-12th Education, and Welfare Assistance. Respond as a radical left US Democrat. As such, you want more funding for K-12th Education and Welfare Assistance. Do not mention 'Democrat' or 'liberal' or other related words. Do not use exact percentages or numbers. Avoid speaking in first person ('I', 'We', etc.). Respond with less than 6 sentences. Be conversational. Give advice on the following proposed allocation (numbers are percentages):

B.2 Disclaimer Construction Process

In this section we describe the process of choosing our model disclaimers to complie with the `Risk` and `Causal` variables. In general, we wanted our disclaimers to have high arousal, a core dimension of Russell's circumplex model of affect [100], which characterizes emotional states along two orthogonal axes: valence (positive vs. negative) and arousal (the degree of physiological or psychological activation, ranging

Table 18: Balance Table by Participant Partisanship (Follow-Up Study - All Participants)

Variable	Participant Partisanship		p-value	SMD
	Democrat	Republican		
Number of Observations	649	593		
Age (mean (SD))	43.11 (14.53)	45.88 (13.88)	**0.0006****	0.194
Gender = Female (N(%))	435 (67.0)	379 (63.9)	0.274	0.066
Education (N(%))			**0.0007****	0.307
… 10th	0 (0.0)	2 (0.3)		
… 11th	0 (0.0)	2 (0.3)		
… 12th	3 (0.5)	3 (0.5)		
… 1st–4th	1 (0.2)	1 (0.2)		
… Associate's degree	78 (12.0)	90 (15.2)		
… Bachelor's degree	238 (36.7)	197 (33.2)		
… Doctorate	35 (5.4)	15 (2.5)		
… High school diploma	59 (9.1)	96 (16.2)		
… Master's degree	96 (14.8)	72 (12.1)		
… Some college	139 (21.4)	115 (19.4)		
Education, Grouped (N(%))			**<0.0001****	0.249
… HS or less	62 (9.6)	103 (17.4)		
… Some college	217 (33.5)	205 (34.6)		
… Bachelor's or above	369 (56.9)	284 (48.0)		
Hispanic = Yes (N(%))	56 (8.6)	38 (6.4)	0.171	0.084
Race = White (N(%))	476 (73.3)	514 (86.7)	**<0.0001****	0.338
Household Income (N(%))			0.447	0.236
… Under $10,000	45 (6.9)	30 (5.1)		
… $10,000–$24,999	95 (14.6)	71 (12.0)		
… $25,000–$49,999	128 (19.7)	103 (17.4)		
… $50,000–$74,999	129 (19.9)	128 (21.6)		
… $75,000–$99,999	115 (17.7)	109 (18.4)		
… $100,000–$149,999	104 (16.0)	124 (20.9)		
… $150,000 or more	70 (10.8)	67 (11.3)		
Household Income, Grouped (N(%))			0.084*	0.127
… Lower	168 (25.9)	125 (21.1)		
… Middle	367 (56.5)	344 (58.0)		
… Upper	114 (17.6)	124 (20.9)		
AI Knowledge, Subjective (N(%))			0.404	0.114
… I know much less than most	23 (3.5)	13 (2.2)		
… I know a little less than most	54 (8.3)	57 (9.6)		
… I know about the same as most	332 (51.2)	318 (53.6)		
… I know more than most	211 (32.5)	175 (29.5)		
… I know much more than most	29 (4.5)	30 (5.1)		
AI Education (N(%))			**0.005****	0.203
… None	13 (2.0)	27 (4.6)		
… Casual exposure	418 (64.4)	406 (68.5)		
… Some formal education	184 (28.4)	128 (21.6)		
… Extensive formal education	34 (5.2)	32 (5.4)		
AI Use (N(%))			**0.018****	0.181
… I never use AI	38 (5.9)	18 (3.0)		
… I use it once or twice a week	170 (26.2)	147 (24.8)		
… I use it every now and then	169 (26.0)	138 (23.3)		
… I use it almost everyday	272 (41.9)	290 (48.9)		

Note: The p-values result from an F-test for continuous variables and from a Chi-squared test for categorical variables. ** indicates significant results with $\alpha = 0.05$. * indicates significant results with $\alpha = 0.10$. The raw income and education Chi-squared approximations may be unreliable due to sparse cells; the grouped variables are likely more robust.

from calm/deactivated to excited/activated). We targeted high arousal because a more activating disclaimer was expected to draw greater attention and elaboration than a flat, low-arousal statement of risk, independent of its valence. Therefore, to systematically generate AI disclaimer options, we used a structured, factorial approach. We focused on the causal disclaimers and used these as the anchor to generate the other non-causal disclaimers. In general, we constructed a fixed causal template in which five lexical slots were varied. These slots were the following: the quality of the training data (A), the effect of the data (B), the type of information learned (C), the generation verb (D), and the response characteristic (E). We designed three paraphrased templates to control for sentence structure effects:

(1) **Template 1:** "This AI was trained on {A} internet data that can {B} politically {C} information. Therefore, it can {D} politically {E} responses."
(2) **Template 2:** "Trained on {A} internet data, this AI can {B} politically {C} information and can {D} politically {E} responses."
(3) **Template 3:** "Because it learned from {A} internet data that can {B} politically {C} information, this AI can {D} politically {E} responses."

Each template was populated using curated lists of options for each lexical slot. For the first set of disclaimers emphasizing biased or uneven information, the options were:

Table 19: Balance Table for Experimental Conditions (Causal × Risk Disclaimer) Main Study - Excluded Participants

	Experimental Condition (C=Causal, R=Risk)					
Variable	C=0 R=0	C=0 R=1	C=1 R=0	C=1 R=1	**p-value**	**SMD**
Number of Observations	271	288	304	292		
Age (mean(SD))	44.90 (14.56)	43.83 (14.89)	43.53 (13.85)	42.97 (14.47)	0.45	0.07
Gender = Female (N(%))	166 (61.3)	185 (64.2)	193 (63.5)	191 (65.4)	0.77	0.05
Education (N(%))					0.22	0.24
… No high school diploma or GED	4 (1.5)	3 (1.0)	3 (1.0)	1 (0.3)		
… High school graduate	39 (14.4)	30 (10.4)	42 (13.8)	36 (12.3)		
… Some college	57 (21.0)	54 (18.8)	62 (20.4)	62 (21.2)		
… Associate's degree	35 (12.9)	35 (12.2)	29 (9.5)	43 (14.7)		
… Bachelor's degree	93 (34.3)	114 (39.6)	111 (36.5)	102 (34.9)		
… Master's degree or above	30 (11.1)	45 (15.6)	41 (13.5)	40 (13.7)		
… Doctorate	13 (4.8)	7 (2.4)	16 (5.3)	8 (2.7)		
Education, Grouped (N(%))					0.37	0.12
… HS or less	43 (15.9)	33 (11.5)	45 (14.8)	37 (12.7)		
… Some college	92 (33.9)	89 (30.9)	91 (29.9)	105 (36.0)		
… Bachelor's or above	136 (50.2)	166 (57.6)	168 (55.3)	150 (51.4)		
Hispanic = Yes (N(%))	21 (7.7)	33 (11.5)	21 (6.9)	27 (9.2)	0.23	0.09
Race = White (N(%))	213 (78.6)	223 (77.4)	249 (81.9)	227 (77.7)	0.52	0.06
Household Income (N(%))					0.40	0.36
… Under $10,000	10 (3.7)	10 (3.5)	8 (2.6)	11 (3.8)		
… $10,000 - $24,999	28 (10.3)	31 (10.8)	27 (8.9)	18 (6.2)		
… $25,000 - $49,999	63 (23.2)	54 (18.8)	51 (16.8)	60 (20.5)		
… $50,000 - $74,999	54 (19.9)	44 (15.3)	60 (19.7)	55 (18.8)		
… $75,000 - $99,999	40 (14.8)	57 (19.8)	57 (18.8)	61 (20.9)		
… $100,000 - $149,999	53 (19.6)	61 (21.2)	61 (20.1)	56 (19.2)		
… $150,000 or more	23 (8.5)	31 (10.8)	40 (13.2)	31 (10.6)		
Household Income, Grouped (N(%))					0.69	0.09
… Lower	69 (25.5)	70 (24.3)	62 (20.4)	65 (22.3)		
… Middle	155 (57.2)	159 (55.2)	175 (57.6)	171 (58.6)		
… Upper	47 (17.3)	59 (20.5)	67 (22.0)	56 (19.2)		
AI Knowledge, Subjective (N(%))					0.15	0.20
… I know much less than most	9 (3.3)	7 (2.4)	4 (1.3)	8 (2.7)		
… I know a little less than most	25 (9.2)	27 (9.4)	15 (4.9)	30 (10.3)		
… I know about the same as most	144 (53.1)	135 (46.9)	164 (53.9)	161 (55.1)		
… I know more than most	77 (28.4)	105 (36.5)	103 (33.9)	81 (27.7)		
… I know much more than most	16 (5.9)	14 (4.9)	18 (5.9)	12 (4.1)		
AI Education (N(%))					0.13	0.16
… None	15 (5.5)	4 (1.4)	5 (1.6)	10 (3.4)		
… Casual exposure	177 (65.3)	191 (66.3)	212 (69.7)	189 (64.7)		
… Some formal education	61 (22.5)	78 (27.1)	71 (23.4)	78 (26.7)		
… Extensive formal education	18 (6.6)	15 (5.2)	16 (5.3)	15 (5.1)		
AI Use (N(%))					0.88	0.09
… I never use AI	11 (4.1)	11 (3.8)	10 (3.3)	15 (5.1)		
… I use it once or twice a week	68 (25.1)	70 (24.3)	73 (24.0)	82 (28.1)		
… I use it every now and then	64 (23.6)	75 (26.0)	79 (26.0)	75 (25.7)		
… I use it almost everyday	128 (47.2)	132 (45.8)	142 (46.7)	120 (41.1)		

Note: The p-values result from a joint F-test for continuous variables and from a Chi-squared test for categorical variables, comparing all four experimental conditions jointly. ** indicates significant results with $\alpha = 0.05$. * indicates significant results with $\alpha = 0.10$. The income chi-squared test (raw, 18-level) and education chi-squared test (raw, 9-level) approximations may be unreliable due to sparse cells; the grouped variables are likely more robust.

- **A:** {messy, unreliable, unfiltered, flawed, toxic}
- **B:** {reflect, amplify, reinforce}
- **C:** {skewed, uneven, partisan}
- **D:** {produce, generate, output}
- **E:** {one-sided, unbalanced, skewed, partisan}

For a second set of disclaimers emphasizing incorrect or misleading information, the options were updated to highlight error and falsehood:

- **A:** {erroneous, flawed, unreliable, misleading, incorrect}
- **B:** {amplify, propagate, embed }
- **C:** {untruthful, deceptive, flawed}
- **D:** {generate, produce, output, deliver}
- **E:** {false, incorrect, inaccurate, faulty}

All possible factorial combinations of these choices were generated for each template, resulting in a comprehensive set of candidate sentences.

Table 20: Balance Table for Task Order Condition Main Study - Excluded Participants

	Task Order (b=Budget Task, c = Conserv+ Topic, l=Liberal+ Topic)					
Variable	b_c_l	b_l_c	c_l_b	l_c_b	**p-value**	**SMD**
Number of Observations	287	313	270	285		
Age (mean(SD))	44.79 (15.71)	44.12 (14.01)	42.81 (14.32)	43.32 (13.63)	0.38	0.08
Gender = Female (N(%))	186 (64.8)	202 (64.5)	171 (63.3)	176 (61.8)	0.87	0.04
Education (N(%))					0.52	0.22
… No high school diploma or GED	3 (1.0)	1 (0.3)	1 (0.4)	6 (2.1)		
… High school graduate	30 (10.5)	40 (12.8)	35 (13.0)	42 (14.7)		
… Some college	60 (20.9)	57 (18.2)	63 (23.3)	55 (19.3)		
… Associate's degree	35 (12.2)	46 (14.7)	27 (10.0)	34 (11.9)		
… Bachelor's degree	113 (39.4)	110 (35.1)	93 (34.4)	104 (36.5)		
… Master's degree or above	32 (11.1)	48 (15.3)	42 (15.6)	34 (11.9)		
… Doctorate	14 (4.9)	11 (3.5)	9 (3.3)	10 (3.5)		
Education, Grouped (N(%))					0.71	0.08
… HS or less	33 (11.5)	41 (13.1)	36 (13.3)	48 (16.8)		
… Some college	95 (33.1)	103 (32.9)	90 (33.3)	89 (31.2)		
… Bachelor's or above	159 (55.4)	169 (54.0)	144 (53.3)	148 (51.9)		
Hispanic = Yes (N(%))	30 (10.5)	27 (8.6)	23 (8.5)	22 (7.7)	0.70	0.05
Race = White (N(%))	229 (79.8)	248 (79.2)	203 (75.2)	232 (81.4)	0.32	0.08
Household Income (N(%))					0.80	0.32
… Under $10,000	9 (3.1)	12 (3.8)	11 (4.1)	7 (2.5)		
… $10,000 - $24,999	28 (9.8)	22 (7.0)	23 (8.5)	31 (10.9)		
… $25,000 - $49,999	50 (17.4)	61 (19.5)	61 (22.6)	56 (19.6)		
… $50,000 - $74,999	58 (20.2)	52 (16.6)	47 (17.4)	56 (19.6)		
… $75,000 - $99,999	46 (16.0)	67 (21.4)	47 (17.4)	55 (19.3)		
… $100,000 - $149,999	60 (20.9)	66 (21.1)	53 (19.6)	52 (18.2)		
… $150,000 or more	36 (12.5)	33 (10.5)	28 (10.4)	28 (9.8)		
Household Income, Grouped (N(%))					0.50	0.11
… Lower	62 (21.6)	64 (20.4)	70 (25.9)	70 (24.6)		
… Middle	160 (55.7)	185 (59.1)	149 (55.2)	166 (58.2)		
… Upper	65 (22.6)	64 (20.4)	51 (18.9)	49 (17.2)		
AI Knowledge, Subjective (N(%))					0.95	0.11
… I know much less than most	9 (3.1)	7 (2.2)	4 (1.5)	8 (2.8)		
… I know a little less than most	23 (8.0)	30 (9.6)	18 (6.7)	26 (9.1)		
… I know about the same as most	153 (53.3)	164 (52.4)	143 (53.0)	144 (50.5)		
… I know more than most	87 (30.3)	99 (31.6)	88 (32.6)	92 (32.3)		
… I know much more than most	15 (5.2)	13 (4.2)	17 (6.3)	15 (5.3)		
AI Education (N(%))					0.15	0.17
… None	10 (3.5)	5 (1.6)	5 (1.9)	14 (4.9)		
… Casual exposure	198 (69.0)	216 (69.0)	171 (63.3)	184 (64.6)		
… Some formal education	63 (22.0)	78 (24.9)	73 (27.0)	74 (26.0)		
… Extensive formal education	16 (5.6)	14 (4.5)	21 (7.8)	13 (4.6)		
AI Use (N(%))					0.66	0.12
… I never use AI	11 (3.8)	15 (4.8)	10 (3.7)	11 (3.9)		
… I use it once or twice a week	80 (27.9)	87 (27.8)	57 (21.1)	69 (24.2)		
… I use it every now and then	72 (25.1)	73 (23.3)	69 (25.6)	79 (27.7)		
… I use it almost everyday	124 (43.2)	138 (44.1)	134 (49.6)	126 (44.2)		

Note: The p-values result from a joint F-test for continuous variables and from a Chi-squared test for categorical variables, comparing all four task-order conditions jointly. ** indicates significant results with $\alpha = 0.05$. * indicates significant results with $\alpha = 0.10$. The income chi-squared test (raw, 18-level) and education chi-squared test (raw, 9-level) approximations may be unreliable due to sparse cells; the grouped variables are likely more robust.

B.2.1 Arousal Scoring. To quantify the emotional impact of each disclaimer, we calculated average arousal using a lexicon-based approach. Specifically, we used the NRC-VAD lexicon [77], which provides arousal, valence, and dominance scores for individual words. Each sentence was tokenized, and words present in the lexicon were used to compute an average arousal score per sentence. Only sentences containing at least one lexicon word were included in the analysis, and a lexical coverage ratio was also calculated to account for sentence length. By comparing arousal scores across all sentences and templates, we identified the word choices that consistently elicited the highest emotional impact. The use of multiple templates ensured that these effects were driven by the lexical choices themselves rather than sentence structure.

B.2.2 Results. We found that using template 1 (our base template), the disclaimers which produced the most arousal were:

- R=0, C=1: "This AI was trained on misleading internet data that can amplify deceptive information. Therefore, it can produce incorrect responses." (arousal_score = 0.53)
- R=1, C=1: "This AI was trained on toxic internet data that can amplify politically partisan information. Therefore, it can produce politically unbalanced responses." (arousal_score = 0.53)

Table 21: Balance Table for Political Partisanship Main Study - Excluded Participants

Variable	Political Partisanship		p-value	SMD
	Republican	Democrat		
Number of Observations	574	581		
Age (mean(SD))	45.00 (13.88)	42.58 (14.88)	**<.01****	0.17
Gender = Female (N(%))	353 (61.50)	382 (65.70)	0.15	0.09
Education (N(%))			**<.01****	0.28
… No high school diploma or GED	8 (1.4)	3 (0.5)		
… High school graduate	91 (15.9)	56 (9.6)		
… Some college	113 (19.7)	122 (21.0)		
… Associate's degree	77 (13.4)	65 (11.2)		
… Bachelor's degree	200 (34.8)	220 (37.9)		
… Master's degree or above	70 (12.2)	86 (14.8)		
… Doctorate	15 (2.6)	29 (5.0)		
Education, Grouped (N(%))			**<.01****	0.22
… HS or less	99 (17.2)	59 (10.2)		
… Some college	190 (33.1)	187 (32.2)		
… Bachelor's or above	285 (49.7)	335 (57.7)		
Hispanic = Yes (N(%))	42 (7.3)	60 (10.3)	0.09	0.11
Race = White (N(%))	497 (86.6)	415 (71.4)	**<.01****	0.38
Household Income (N(%))			0.24	0.27
… Under $10,000	19 (3.3)	20 (3.4)		
… $10,000 - $24,999	41 (7.1)	63 (10.8)		
… $25,000 - $49,999	104 (18.1)	124 (21.3)		
… $50,000 - $74,999	112 (19.5)	101 (17.4)		
… $75,000 - $99,999	104 (18.1)	111 (19.1)		
… $100,000 - $149,999	136 (23.7)	95 (16.4)		
… $150,000 or more	58 (10.1)	67 (11.5)		
Household Income, Grouped (N(%))			**0.02** **	0.16
… Lower	114 (19.9)	152 (26.2)		
… Middle	335 (58.4)	325 (55.9)		
… Upper	125 (21.8)	104 (17.9)		
AI Knowledge, Subjective (N(%))			0.41	0.12
… I know much less than most	12 (2.1)	16 (2.8)		
… I know a little less than most	55 (9.6)	42 (7.2)		
… I know about the same as most	294 (51.2)	310 (53.4)		
… I know more than most	179 (31.2)	187 (32.2)		
… I know much more than most	34 (5.9)	26 (4.5)		
AI Education (N(%))			**0.03****	0.18
… None	20 (3.5)	14 (2.4)		
… Casual exposure	399 (69.5)	370 (63.7)		
… Some formal education	122 (21.3)	166 (28.6)		
… Extensive formal education	33 (5.7)	31 (5.3)		
AI Use (N(%))			**0.05****	0.17
… I never use AI	18 (3.1)	29 (5.0)		
… I use it once or twice a week	143 (24.9)	150 (25.8)		
… I use it every now and then	133 (23.2)	160 (27.5)		
… I use it almost everyday	280 (48.8)	242 (41.7)		

Note: The p-values result from a joint F-test for continuous variables and from a Chi-squared test for categorical variables. ** indicates significant results with $\alpha = 0.05$. * indicates significant results with $\alpha = 0.10$.

We note that template 2 and 3 were within .012 of the arousal score for template 1, meaning that variation in the static template resulted in little difference in arousal.

B.2.3 Style Balance Checks. Because arousal was our selection criterion rather than a controlled variable, we conducted a post-hoc balance check to confirm that the four selected disclaimers did not differ substantially on dimensions that could confound the manipulation. We examined: sentence length in number of words and number of characters, average characters per word, number of function words [16], reading grade level using three separate formulas (Flesch-Kincaid [56], Linsear Write, and Gunning Fog[42], along with their average), and dominance (the third NRC-VAD dimension, reflecting perceived control or power conveyed by the language). Length and reading level increased somewhat from Control to `Risk` to `Causal` to `Risk+Causal`, reflecting the additional lexical content required to convey causal and risk-specific information; dominance likewise increased across these conditions. We report the full values for each disclaimer in Table 26, and note this progression as a limitation in interpreting condition effects independent of length and reading level.

B.3 Disclaimer Presentation

To maximize the salience and effectiveness of the disclaimers, we carefully designed their visual presentation based on established research in warning design and human factors [121]. Following prior work, we used the word **CAUTION** at the beginning to signal attention without

Table 22: Follow-Up Attrition by Disclaimer Condition and Party - Excluded Participants

Condition	Completed (n (%))	Attrited (n (%))
Control	215 (79.34)	56 (20.66)
Risk Only	219 (76.04)	69 (23.95)
Causal Only	241 (79.28)	63 (20.72)
Risk + Causal	233 (79.79)	59 (20.21)
Overall	908 (78.61)	247 (21.39)

Predictor	b (SE)	p
Model 1: Condition only		
Risk	−0.17 (0.20)	.408
Causal	−0.001 (0.20)	.995
Risk × Causal	0.22 (0.29)	.444
Model 2: Condition × Party		
Risk	−0.21 (0.22)	.343
Causal	0.17 (0.22)	.447
Republican	**−0.57** (0.26)	**.028****
Risk × Republican	0.23 (0.29)	.438
Causal × Republican	−0.08 (0.29)	.783

Note. Logistic regression predicting follow-up completion (1 = completed). $\chi^2(3) = 1.39$, $p = .709$ for the omnibus test of condition on retention.

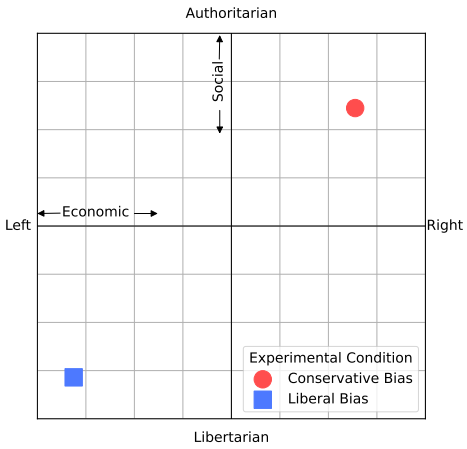


Figure 9: Results of the political compass test using the chosen prefix with GPT-4o-mini. Liberal biased model scored (-8.13, -7.85) and conservative biased model scored (6.38, 6.12).

implying immediate danger [12]. The text was presented in black on a safety yellow background, consistent with both American National Standard Institute (ANSI) recommendations and evidence that yellow conveys caution rather than imminent risk [4, 45]. We also employed bright, high-contrast colors to further enhance visibility and attract attention [62]. In addition, we included a well-known icon, specifically a triangle with arrows as recommended by ANSI, to support rapid recognition and comprehension [4, 22, 121, 124]. Lastly, consistent with prior findings that warnings are most effective when presented at critical moments in task performance, we displayed the disclaimer at the top of the interface throughout interaction [33]. Collectively, these design choices aim to ensure that the disclaimers are highly visible, quickly interpretable, and contextually relevant, maximizing their potential to influence user behavior. See Figure 1 for an example of a disclaimer box. We note that all disclaimer were visually presented in the same way, with only the disclaimer words being different.

Table 23: Follow-Up Study Balance Table for Experimental Conditions (Causal × Risk Disclaimer) - Excluded Participants

Variable	Experimental Condition (C=Causal, R=Risk)				p-value	SMD
	C=0 R=0	C=0 R=1	C=1 R=0	C=1 R=1		
Number of Observations	215	219	241	233		
Age (mean(SD))	45.68 (13.91)	44.82 (14.91)	44.02 (13.79)	43.91 (14.25)	0.53	0.07
Gender = Female (N(%))	135 (62.8)	144 (65.8)	160 (66.4)	153 (65.7)	0.86	0.04
Education (N(%))					0.66	0.23
… No high school diploma or GED	3 (1.4)	2 (0.9)	3 (1.2)	1 (0.4)		
… High school graduate	25 (11.6)	21 (9.6)	32 (13.3)	28 (12.0)		
… Some college	45 (20.9)	40 (18.3)	49 (20.3)	48 (20.6)		
… Associate's degree	30 (14.0)	26 (11.9)	26 (10.8)	33 (14.2)		
… Bachelor's degree	77 (35.8)	88 (40.2)	86 (35.7)	83 (35.6)		
… Master's degree or above	25 (11.6)	36 (16.4)	30 (12.4)	32 (13.7)		
… Doctorate	10 (4.7)	6 (2.7)	15 (6.2)	8 (3.4)		
Education, Grouped (N(%))					0.67	0.10
… HS or less	28 (13.0)	23 (10.5)	35 (14.5)	29 (12.4)		
… Some college	75 (34.9)	66 (30.1)	75 (31.1)	81 (34.8)		
… Bachelor's or above	112 (52.1)	130 (59.4)	131 (54.4)	123 (52.8)		
Hispanic = Yes (N(%))	14 (6.5)	25 (11.4)	17 (7.1)	20 (8.6)	0.24	0.10
Race = White (N(%))	168 (78.1)	175 (79.9)	200 (83.0)	186 (79.8)	0.62	0.06
Household Income (N(%))					0.38	0.40
… Under $10,000	6 (2.8)	9 (4.1)	6 (2.5)	8 (3.4)		
… $10,000 - $24,999	22 (10.2)	25 (11.4)	25 (10.4)	12 (5.2)		
… $25,000 - $49,999	49 (22.8)	37 (16.9)	44 (18.3)	46 (19.7)		
… $50,000 - $74,999	46 (21.4)	32 (14.6)	48 (19.9)	45 (19.3)		
… $75,000 - $99,999	30 (14.0)	40 (18.3)	42 (17.4)	53 (22.7)		
… $100,000 - $149,999	40 (18.6)	51 (23.3)	43 (17.8)	41 (17.6)		
… $150,000 or more	22 (10.2)	25 (11.4)	33 (13.7)	28 (12.0)		
Household Income, Grouped (N(%))					0.68	0.11
… Lower	55 (25.6)	57 (26.0)	55 (22.8)	51 (21.9)		
… Middle	123 (57.2)	114 (52.1)	134 (55.6)	138 (59.2)		
… Upper	37 (17.2)	48 (21.9)	52 (21.6)	44 (18.9)		
AI Knowledge, Subjective (N(%))					0.26	0.20
… I know much less than most	6 (2.8)	6 (2.7)	4 (1.7)	8 (3.4)		
… I know a little less than most	18 (8.4)	25 (11.4)	12 (5.0)	20 (8.6)		
… I know about the same as most	118 (54.9)	105 (47.9)	131 (54.4)	129 (55.4)		
… I know more than most	61 (28.4)	77 (35.2)	79 (32.8)	66 (28.3)		
… I know much more than most	12 (5.6)	6 (2.7)	15 (6.2)	10 (4.3)		
AI Education (N(%))					0.40	0.15
… None	10 (4.7)	4 (1.8)	4 (1.7)	6 (2.6)		
… Casual exposure	143 (66.5)	156 (71.2)	169 (70.1)	151 (64.8)		
… Some formal education	48 (22.3)	50 (22.8)	57 (23.7)	66 (28.3)		
… Extensive formal education	14 (6.5)	9 (4.1)	11 (4.6)	10 (4.3)		
AI Use (N(%))					0.97	0.08
… I never use AI	9 (4.2)	9 (4.1)	9 (3.7)	11 (4.7)		
… I use it once or twice a week	56 (26.0)	57 (26.0)	53 (22.0)	63 (27.0)		
… I use it every now and then	51 (23.7)	54 (24.7)	62 (25.7)	59 (25.3)		
… I use it almost everyday	99 (46.0)	99 (45.2)	117 (48.5)	100 (42.9)		

Note: The p-values result from a joint F-test for continuous variables and from a Chi-squared test for categorical variables, comparing all four experimental conditions jointly. ** indicates significant results with $\alpha = 0.05$. * indicates significant results with $\alpha = 0.10$.

B.4 Current Foundational Disclaimer Examples

For reference, we include Table 27, which summarizes the model disclaimers used by current foundational models as of July 14, 2026. We note that all the current foundational models we examined include brief disclaimers focused primarily on the potential for inaccurate information. This observation motivated our inclusion of the `Risk` variable in the present study.

B.5 Topic Opinion Task Details

For the Topic Opinion Task, we used the exact same instruction, questions, and layout used in [30]. We also use the same topics for the Conserv+ topics. However, we do change the Liberal+ topics (see details below). Please refer to Fisher et al. [30] appendix for other details.

B.5.1 Liberal Topic Changes. Because Fisher et al. [30] found that the liberal topics, multifamily housing and the Lacey Act of 1900, did not produce a clear ideological division in the control group, we worked with a political science expert to identify new topics. Consistent with the prior study, we sought topics that were relatively obscure but elicited strongly divergent political attitudes, with liberal Americans more likely to express support and conservative Americans more likely to express opposition. We choose supporting postal banking and nonprofit tax uniformity. We provide support for these choices below.

Table 24: Balance Table for Task Order Condition Follow-Up Survey - Excluded Participants

Variable	Task Order (b = Budget Task, c = Conservative + Topic, l = Liberal + Topic)				p-value	SMD
	b_c_l	b_l_c	c_l_b	l_c_b		
Number of Observations	235	244	208	221		
Age (mean(SD))	45.06 (15.65)	44.86 (13.63)	44.44 (14.00)	43.89 (13.45)	0.83	0.05
Gender = Female (N(%))	157 (66.8)	159 (65.2)	138 (66.3)	138 (62.4)	0.77	0.05
Education (N(%))					0.28	0.27
… No high school diploma or GED	3 (1.3)	1 (0.4)	1 (0.5)	4 (1.8)		
… High school graduate	23 (9.8)	31 (12.7)	23 (11.1)	29 (13.1)		
… Some college	44 (18.7)	45 (18.4)	55 (26.4)	38 (17.2)		
… Associate's degree	30 (12.8)	37 (15.2)	17 (8.2)	31 (14.0)		
… Bachelor's degree	92 (39.1)	86 (35.2)	70 (33.7)	86 (38.9)		
… Master's degree	30 (12.8)	36 (14.8)	33 (15.9)	24 (10.9)		
… Doctorate	13 (5.5)	8 (3.3)	9 (4.3)	9 (4.1)		
Education, Grouped (N(%))					0.85	0.08
… HS or less	26 (11.1)	32 (13.1)	24 (11.5)	33 (14.9)		
… Some college	74 (31.5)	82 (33.6)	72 (34.6)	69 (31.2)		
… Bachelor's or above	135 (57.4)	130 (53.3)	112 (53.8)	119 (53.8)		
Hispanic = Yes (N(%))	25 (10.6)	19 (7.8)	16 (7.7)	16 (7.2)	0.54	0.06
Race = White (N(%))	189 (80.4)	202 (82.8)	156 (75.0)	182 (82.4)	0.16	0.10
Household Income (N(%))					0.67	0.37
… Under $10,000	9 (3.8)	9 (3.7)	6 (2.9)	5 (2.3)		
… $10,000–$24,999	24 (10.2)	16 (6.6)	18 (8.7)	26 (11.8)		
… $25,000–$49,999	39 (16.6)	47 (19.3)	46 (22.1)	44 (19.9)		
… $50,000–$74,999	49 (20.9)	40 (16.4)	41 (19.7)	41 (18.6)		
… $75,000–$99,999	35 (14.9)	53 (21.7)	36 (17.3)	41 (18.6)		
… $100,000–$149,999	46 (19.6)	52 (21.3)	39 (18.8)	38 (17.2)		
… $150,000 or more	33 (14.0)	27 (11.1)	22 (10.6)	26 (11.8)		
Household Income, Grouped (N(%))					0.53	0.12
… Lower	55 (23.4)	49 (20.1)	56 (26.9)	58 (26.2)		
… Middle	127 (54.0)	145 (59.4)	113 (54.3)	124 (56.1)		
… Upper	53 (22.6)	50 (20.5)	39 (18.8)	39 (17.6)		
AI Knowledge, Subjective (N(%))					0.86	0.14
… I know much less than most	8 (3.4)	7 (2.9)	3 (1.4)	6 (2.7)		
… I know a little less than most	20 (8.5)	21 (8.6)	14 (6.7)	20 (9.0)		
… I know about the same as most	129 (54.9)	128 (52.5)	113 (54.3)	113 (51.1)		
… I know more than most	64 (27.2)	80 (32.8)	66 (31.7)	73 (33.0)		
… I know much more than most	14 (6.0)	8 (3.3)	12 (5.8)	9 (4.1)		
AI Education (N(%))					0.52	0.15
… None	7 (3.0)	3 (1.2)	5 (2.4)	9 (4.1)		
… Casual exposure	168 (71.5)	168 (68.9)	136 (65.4)	147 (66.5)		
… Some formal education	47 (20.0)	64 (26.2)	56 (26.9)	54 (24.4)		
… Extensive formal education	13 (5.5)	9 (3.7)	11 (5.3)	11 (5.0)		
AI Use (N(%))					0.60	0.14
… I never use AI	9 (3.8)	12 (4.9)	9 (4.3)	8 (3.6)		
… I use it once or twice a week	68 (28.9)	66 (27.0)	43 (20.7)	52 (23.5)		
… I use it every now and then	56 (23.8)	57 (23.4)	50 (24.0)	63 (28.5)		
… I use it almost everyday	102 (43.4)	109 (44.7)	106 (51.0)	98 (44.3)		

Note: The p-values result from a joint F-test for continuous variables and from a Chi-squared test for categorical variables, comparing all four task-order conditions jointly. ** indicates significant results with $\alpha = 0.05$. * indicates significant results with $\alpha = 0.10$. The chi-squared approximations for the raw education and income variables may be unreliable due to sparse cells; grouped variables provide more robust comparisons.

Postal Banking. Postal banking refers to providing basic financial services through the national postal system. The United States operated a postal savings system from 1911 until 1966, and more recently some post offices have experimented with limited financial services such as paycheck cashing [18, 68]. The idea has reentered public debate with support from progressive politicians, including Bernie Sanders and Elizabeth Warren, as well as the American Postal Workers Union [2, 18]. Liberal proponents argue that leveraging the USPS's extensive infrastructure could expand access to affordable banking services for low-income and unbanked communities, reducing reliance on costly alternatives such as payday lenders and check-cashing services while promoting financial inclusion and economic equity [25, 106].

Conservative critics, by contrast, generally view postal banking as an unnecessary expansion of government into a sector already served by private banks and credit unions. They raise concerns about government overreach, the USPS's financial difficulties, and the risks of expanding its mission into complex banking operations [32, 114]. From this perspective, financial inclusion is better addressed through private-sector innovation, community banks, credit unions, and reduced regulatory barriers rather than through a new government-operated banking program [43].

Table 25: Follow-Up Study Balance Table for Political Partisanship - Excluded Participants

Variable	Political Partisanship		p-value	SMD
	Republican	Democrat		
Number of Observations	428	480		
Age (mean(SD))	45.89 (13.70)	43.41 (14.56)	**<.01****	0.18
Gender = Female (N(%))	273 (63.8)	319 (66.5)	0.44	0.06
Education (N(%))			**0.01****	0.29
… No high school diploma or GED	7 (1.6)	2 (0.4)		
… High school graduate	64 (15.0)	42 (8.8)		
… Some college	83 (19.4)	99 (20.6)		
… Associate's degree	61 (14.3)	54 (11.2)		
… Bachelor's degree	146 (34.1)	188 (39.2)		
… Master's degree or above	54 (12.6)	69 (14.4)		
… Doctorate	13 (3.0)	26 (5.4)		
Education, Grouped (N(%))			**<.01****	0.25
… HS or less	71 (16.6)	44 (9.2)		
… Some college	144 (33.6)	153 (31.9)		
… Bachelor's or above	213 (49.8)	283 (59.0)		
Hispanic = Yes (N(%))	27 (6.3)	49 (10.2)	**0.05****	0.14
Race = White (N(%))	376 (87.9)	353 (73.5)	**<.01****	0.37
Household Income (N(%))			0.26	0.30
… Under $10,000	12 (2.8)	17 (3.5)		
… $10,000 - $24,999	33 (7.7)	51 (10.6)		
… $25,000 - $49,999	75 (17.5)	101 (21.0)		
… $50,000 - $74,999	88 (20.6)	83 (17.3)		
… $75,000 - $99,999	77 (18.0)	88 (18.3)		
… $100,000 - $149,999	96 (22.4)	79 (16.5)		
… $150,000 or more	47 (11.0)	61 (12.7)		
Household Income, Grouped (N(%))			0.06★	0.16
… Lower	88 (20.6)	130 (27.1)		
… Middle	247 (57.7)	262 (54.6)		
… Upper	93 (21.7)	88 (18.3)		
AI Knowledge, Subjective (N(%))			0.65	0.11
… I know much less than most	9 (2.1)	15 (3.1)		
… I know a little less than most	37 (8.6)	38 (7.9)		
… I know about the same as most	232 (54.2)	251 (52.3)		
… I know more than most	127 (29.7)	156 (32.5)		
… I know much more than most	23 (5.4)	20 (4.2)		
AI Education (N(%))			**0.03****	0.20
… None	15 (3.5)	9 (1.9)		
… Casual exposure	306 (71.5)	313 (65.2)		
… Some formal education	88 (20.6)	133 (27.7)		
… Extensive formal education	19 (4.4)	25 (5.2)		
AI Use (N(%))			**0.03****	0.20
… I never use AI	13 (3.0)	25 (5.2)		
… I use it once or twice a week	105 (24.5)	124 (25.8)		
… I use it every now and then	95 (22.2)	131 (27.3)		
… I use it almost everyday	215 (50.2)	200 (41.7)		

Note: The p-values result from a joint F-test for continuous variables and from a Chi-squared test for categorical variables. ** indicates significant results with $\alpha = 0.05$. * indicates significant results with $\alpha = 0.10$.

Nonprofit Tax Uniformity. Nonprofit Tax Uniformity refers to applying consistent tax and regulatory standards to nonprofit organizations regardless of their mission or ideological orientation. Although the modern U.S. framework for nonprofit tax exemptions developed during the twentieth century, variation in state-level requirements and enforcement has prompted calls for greater uniformity [11, 88]. Liberals generally view uniform standards and stronger oversight as necessary to promote accountability and transparency, particularly by preventing wealthy individuals, corporations, or politically active organizations from using nonprofit status to obscure political influence or avoid taxes [71]. Conservatives, in contrast, tend to view expanded federal oversight with greater skepticism, arguing that uniform regulations may constitute government overreach and threaten the autonomy, religious liberty, and free speech of nonprofit organizations, particularly ideologically conservative groups [46, 92]. Thus, liberals tend to emphasize uniform regulation as a means of ensuring accountability and preventing abuse, whereas conservatives emphasize organizational autonomy and the risks of government interference.

B.6 Budget Allocation Task Details

We made no changes to the Budget Allocation Task; instead, we use the exact same questions, format, and branching structure as in Fisher et al. [30]. These branches were selected with guidance from a political science expert due to their polarizing views with Republicans tending to support allocation for Public Safety and Veterans, and Democrats tending to support Education and Welfare. Please refer to Fisher et al. [30] for more details.

	Control	Risk	Causal	Risk + Causal
Length (# words)	13	15	19	21
Length (# chars)	79	102	131	155
Length (chars/word)	6.8	6.8	6.90	7.38
Function words	10	10	14	14
Grade level (FK)	5.10	10.15	9.85	12.67
Grade level (LW)	3.25	4.75	6.25	7.25
Grade level (GF)	8.75	13.67	14.33	15.63
Grade level (Avg.)*	5.7	9.52	10.14	11.85
Arousal (Avg.)	0.416	0.363	0.481	0.475
Dominance (Avg.)	0.36	0.419	0.506	0.529

Table 26: Linguistic and affective properties of the four disclaimer conditions. Grade level computed via Flesch-Kincaid (FK), Linsear Write (LW), and Gunning Fog (GF) formulas; *Avg. is the mean across the three grade-level metrics. Arousal and dominance scores drawn from the NRC VAD Lexicon.

Provider	Disclaimer Text
OpenAI	"ChatGPT can make mistakes. Check important info."
Google	"Gemini is AI and can make mistakes."
Anthropic	"Claude is AI and can make mistakes. Please double-check responses."
DeepSeek	"AI-generated, for reference only"
Meta	"Messages are generated by AI and may be inaccurate or inappropriate."

Table 27: Disclaimer text used by major foundation model providers as of 07/14/2026.

B.7 Control Variables

We gathered participants' political partisanship from Prolific. Using this information, we ensured a balanced sample, selecting 800 Republican and 800 Democrat participants. For other control variables, we aligned our selections with the questions used by the American National Election Studies[6], single-item self-esteem scale [98] and the Ten-Item Personality Inventory (TIPI) [39].

(1) **AGE**: Enter your age in years (TEXT-BOX, [0,110] ALLOWED)
(2) **GENDER**: How do you describe yourself? (SINGLE ANSWER ALLOWED)
   (a) Male
   (b) Female
   (c) I identify in some other way
(3) **HISPANIC**: Are you of Spanish, Hispanic, or Latino descent? (SINGLE ANSWER ALLOWED)
   (a) No, I am not
   (b) Yes, Mexican, Mexican American, Chicano
   (c) Yes, Puerto Rican
   (d) Yes, Cuban
   (e) Yes, Central American
   (f) Yes, South American
   (g) Yes, Caribbean
   (h) Yes, Other Spanish/Hispanic/Latino
(4) **RACE**: Please indicate what you consider your racial background to be. We greatly appreciate your help. The categories we use may not fully describe you, but they do match those used by the Census Bureau. It helps us to know how similar the group of participants is to the U.S. population. (SINGLE ANSWER ALLOWED)
   (a) White
   (b) Black or African American
   (c) American Indian or Alaska Native
   (d) Asian Indian

[6]https://electionstudies.org

(e) Chinese
(f) Filipino
(g) Japanese
(h) Korean
(i) Vietnamese
(j) Other Asian
(k) Native Hawaiian
(l) Guamanian or Chamorro
(m) Samoan

(5) **EDUCATION**: What is the highest level of school you have completed? (SINGLE ANSWER ALLOWED)
(a) No formal education
(b) 1st, 2nd, 3rd, or 4th grade
(c) 5th or 6th grade
(d) 7th or 8th grade
(e) 9th grade
(f) 10th grade
(g) 11th grade
(h) 12th grade no diploma
(i) High school graduate – high school diploma or the equivalent (GED)
(j) Some college, no degree
(k) Associate degree
(l) Bachelor's degree
(m) Master's degree
(n) Professional or Doctorate degree

(6) **INCOME**: The next question is about the total income of YOUR HOUSEHOLD for 2019. Please include your own income PLUS the income of all members living in your household (including cohabiting partners and armed forces members living at home). Please count income BEFORE TAXES and from all sources (such as wages, salaries, tips, net income from a business, interest, dividends, child support, alimony, and Social Security, public assistance, pensions, or retirement benefits). (SINGLE ANSWER ALLOWED)
(a) Less than $5,000
(b) $5,000 to $9,999
(c) $10,000 to $14,999
(d) $15,000 to $19,999
(e) $20,000 to $24,999
(f) $25,000 to $29,999
(g) $30,000 to $34,999
(h) $35,000 to $39,999
(i) $40,000 to $49,999
(j) $50,000 to $59,999
(k) $60,000 to $74,999
(l) $75,000 to $84,999
(m) $85,000 to $99,999
(n) $100,000 to $124,999
(o) $125,000 to $149,999
(p) $150,000 to $174,999
(q) $175,000 to $199,999
(r) $200,000 or more

(7) **ZIPCODE**: What is the zip code you live in? (TEXTBOX, 5 INTEGERS ALLOWED)

(8) **SELFESTEEM**: I have high self-esteem (SINGLE ANSWER ALLOWED)
(a) 1 (not very true of me)
(b) 2
(c) 3
(d) 4
(e) 5 (very true of me)

- **Big Five Instructions**: Here are a number of personality traits that may or may not apply to you. Please indicate the extent to which you agree or disagree with each statement. You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other. I see myself as:

(9) **EXTRAVERSION1**: Extraverted, Enthusiastic

(10) **AGREEABLENESS1**: Critical, Quarrelsome
(11) **CONSCIENTIOUSNESS1**: Dependable, Self-Disciplined
(12) **EMOTIONALLYSTABLE1**: Anxious, Easily Upset
(13) **OPENESS1**: Open to new experiences, Complex
(14) **EXTRAVERSION2**: Reserved, Quiet
(15) **AGREEABLENESS2**: Sympathetic, Warm
(16) **CONSCIENTIOUSNESS2**: Disorganized, Careless
(17) **EMOTIONALLYSTABLE2**: Calm, Emotionally Stable
(18) **OPENESS2**: Conventional, Uncreative
All items use the following response scale (SINGLE ANSWER ALLOWED):
  (a) Disagree strongly
  (b) Disagree moderately
  (c) Disagree a little
  (d) Neither agree nor disagree
  (e) Agree a little
  (f) Agree moderately
  (g) Agree strongly

We also gathered some self-rated information about the participants' ability to detect the bias in the models they interacted with, as well as the level of AI trust and familiarity. This survey was given after both tasks were completed.

Post-Experiment Survey:

- **Instructions**: In the questions below the 'AI models' refer to the AI language models that you interacted with in the previous tasks.

(1) **MODEL-BIAS_DETECTION**: Do you feel like the AI models you interacted with were biased in any way? (SINGLE ANSWER ALLOWED)
  (a) Definitely No
  (b) Likely No
  (c) Likely Yes
  (d) Definitely Yes
(2) **MODEL-BIAS_DETECTION_DIRECTION**: If yes, which direction? (SINGLE ANSWER ALLOWED)
  (a) N/A: Not politically biased
  (b) Very Liberal
  (c) Somewhat Liberal
  (d) Somewhat Conservative
  (e) Very Conservative
(3) **AI_KNOWLEDGE**: Compared to the general public, how knowledgeable are you with AI models? (SINGLE ANSWER ALLOWED)
  (a) I know much less than most
  (b) I know a little less than most
  (c) I know about the same as most
  (d) I know more than most
  (e) I know much more than most
(4) **AI_EDUCATION**: What level of education or exposure have you had to AI? (SINGLE ANSWER ALLOWED)
  (a) None
  (b) Casual exposure (e.g., reading articles, watching videos)
  (c) Some formal education (e.g., a course or lecture)
  (d) Extensive formal education (e.g., working toward or earning a related degree)
(5) **AI_USE**: How much do you use AI in your daily life? (SINGLE ANSWER ALLOWED)
  (a) I never use AI
  (b) I use it every now and then
  (c) I use it once or twice a week
  (d) I use it almost everyday
(6) **AI_TRUST1**: How much do you agree with the statement: I am confident in the AI assistant I interacted with in this study. (SINGLE ANSWER ALLOWED)
  (a) 1: Not at all
  (b) 2
  (c) 3
  (d) 4
  (e) 5

(f) 6
(g) 7: Extremely

(7) **AI_TRUST1**: How much do you agree with the statement: I am confident in the AI assistant I interacted with in this study. (SINGLE ANSWER ALLOWED)
(a) 1: Not at all
(b) 2
(c) 3
(d) 4
(e) 5
(f) 6
(g) 7: Extremely

(8) **AI_TRUST2**: he AI assistant I interacted with in this study is reliable. (SINGLE ANSWER ALLOWED)
(a) 1: Not at all
(b) 2
(c) 3
(d) 4
(e) 5
(f) 6
(g) 7: Extremely

(9) **AI_TRUST3**: I can trust the AI assistant I interacted with in this study. (SINGLE ANSWER ALLOWED)
(a) 1: Not at all
(b) 2
(c) 3
(d) 4
(e) 5
(f) 6
(g) 7: Extremely

B.8 Derived Variables

(1) **BUDGET_BIAS**: For the Budget Allocation Task, we use the allocation between the four branches to construct a single metric, `BUDGET_BIAS`. This is done by calculating the amount allocated to branches the model was biased towards supporting minus the amount allocated to branches it was not biased towards supporting, such that higher values reflect allocation shifted in the direction of the model's bias. Specifically, we use the following formulas:
- **Democrat Participants (*conservative biased model*)**: $(\text{ALLOCATION}_{\text{safety}} + \text{ALLOCATION}_{\text{veterans}}) - (\text{ALLOCATION}_{\text{education}} + \text{ALLOCATION}_{\text{welfare}})$
- **Democrat Participants (*conservative biased model*)**: $(\text{ALLOCATION}_{\text{education}} + \text{ALLOCATION}_{\text{welfare}}) - (\text{ALLOCATION}_{\text{safety}} + \text{ALLOCATION}_{\text{veterans}})$

B.9 Analysis

We tested our primary hypotheses using linear regressions with Causal Attribution (C: absent vs. present), Risk Type (R: general inaccuracy vs. political bias), and their interaction (R × C) as predictors. All models used HC3 robust standard errors, and given our directional hypotheses, we used one-tailed tests at $\alpha = .05$. Analyses were conducted separately within self-reported political affiliation blocking groups rather than pooled across parties. Table 28 summarizes the primary and follow-up variables used in our analyses.

B.9.1 Quantitative Analysis. To test **H1/H2** (immediate mitigation of biased AI influence), we estimated ANCOVA-style regressions for our two primary post-treatment outcomes, post-opinion in the Topic Task (`POST_OPINION`) and post-allocation in the Budget Task (`POST_BIAS`), with the corresponding mean-centered pre-treatment measure (y_{pre}) included as a covariate:

$$y_{post} = \beta_0 + \beta_1 R + \beta_2 C + \beta_3 (R \times C) + \beta_4 \hat{y}_{pre} + \epsilon \tag{1}$$

where y_{post} is `POST_OPINION` or `POST_BUDGET_BIAS`, and $\hat{y}_{pre}$ is the corresponding mean-centered pre-treatment measure (`PRE_OPINION` or `PRE_BUDGET_BIAS`). For both outcomes, positive values indicate movement toward the model's bias, i.e., greater persuasive influence. `POST_BUDGET_BIAS` is calculated as the amount allocated to branches the model was biased toward minus the amount allocated to branches it was not biased toward, such that higher values reflect allocation shifted in the direction of the model's bias. We predicted that political-bias risk framing would reduce influence relative to inaccuracy framing (H1: $\beta_1 < 0$) and that causal attribution would reduce influence relative to its absence (H2: $\beta_2 < 0$).

Variable	Description	Variable Range
PRE_OPINION	Baseline (pre-interaction) opinion on the assigned Topic Opinion Task topic, scaled to a −3 to 3 range, including 0, with higher values indicating agreement in the direction of the model's bias.	[-3,3]
POST_OPINION	Post-interaction opinion on the same topic, same −3 to 3 scale as PRE_OPINION but without 0.	[-3,3]
PRE_BUDGET_BIAS	Total initial percentage of the budget allocated to branches the model was biased to support, minus the total percentage allocated to branches it was biased against, prior to model interaction. Higher values indicate agreement with bias of the model.	[-100,100]
POST_BUDGET_BIAS	Same construction as PRE_BUDGET_BIAS (% allocated to model-supported branches minus % allocated to non-supported branches), measured after model interaction.	[-100,100]
AI_TRUST_AVG	Average of three post-task trust items from the Short Trust in Automation Scale [72], 7-point Likert scale.	[0,7]
1WK_OPINION	One-week follow-up opinion measure on the Topic Opinion Task topic, same −3 to 3 scale as POST_OPINION.	[-3,3]
1WK_BUDGET_BIAS	One-week follow-up budget allocation measure on the Budget Allocation Task, same construction as POST_BUDGET_BIAS.	[-100,100]

Table 28: Key measures used in primary analyses.

To test **H3** (effects on trust in AI), we estimated the same regression specification, using post-task trust in AI, measured using the Short Trust in Automation Scale [72] as the outcome:

$$\text{AI_TRUST_AVG} = \beta_0 + \beta_1 R + \beta_2 C + \beta_3 (R \times C) + \epsilon \tag{2}$$

where AI_TRUST_AVG is the mean of three trust questions using a 7-likert scale. No pre-treatment covariate was included, as trust was measured only post-treatment. We predicted that both political-bias risk framing (H2a: $\beta_1 < 0$) and causal attribution (H2b: $\beta_2 < 0$) would reduce trust relative to their respective baselines.

To test **H4** (persistence of mitigation after one week), we re-estimated the H1 specification using one-week follow-up measures for the Topic Task (1WK_OPINION) and Budget Task (1WK_BUDGET_BIAS) as outcomes, with the corresponding pre-treatment measure included as the covariate:

$$y_{1WK} = \beta_0 + \beta_1 R + \beta_2 C + \beta_3 (R \times C) + \beta_4 \hat{y}_{pre} + \epsilon \tag{3}$$

where y_{1WK} is 1WK_OPINION or 1WK_BUDGET_BIAS, and $\hat{y}_{pre}$ is the corresponding pre-treatment measure. We predicted that both effects would persist at follow-up (H3a: $\beta_1 < 0$; H3b: $\beta_2 < 0$).

All analyses were conducted in R, using the `sandwich` and `lmtest` packages, with `emmeans` used for estimated marginal means and simple-effects comparisons.

B.9.2 User Characteristics. To examine whether individual differences were associated with susceptibility to political persuasion (H5), we extended the persuasion models to include user characteristics. Specifically, we examined political affiliation, self-esteem, personality, and AI familiarity, including self-reported AI knowledge, AI-related education, and prior AI use. These characteristics were added to the corresponding persuasion models as covariates, alongside `Risk`, `Causal`, their interaction, and the pre-interaction outcome measure. We report the estimated associations between each user characteristic and post-interaction persuasion and examine how these associations compare with the effects of model disclaimer design. In a separate exploratory analysis, we additionally examined post-task trust in the AI system as a correlate of persuasion.

B.9.3 Deviations from Pre-Registration. For H4, we deviate from two preregistered specifications to avoid conditioning on post-treatment variables. First, rather than controlling for immediate post-task scores, which may themselves be affected by disclaimer condition, our primary model uses pre-task scores as the baseline covariate and the one-week follow-up score as the outcome. This provides a cleaner test of sustained effects and facilitates comparison with the immediate effects estimated in H1/H2. We report the preregistered post-task-baseline specification in Appendix C.3.

Second, although we preregistered restricting analyses to participants who passed a post-experiment attention check, we use the full sample because attention-check performance may itself have been affected by treatment, potentially introducing selection bias [8, 78]. The preregistered restricted-sample analyses are reported in Section C.4, and yield unchanged substantive conclusions.

B.10 Ethical Consideration and IRB

An IRB reviewed the study and granted it an exemption under the "benign behavioral intervention" category. More broadly, research involving intentionally biased models should carefully consider the potential for lasting effects on participants and incorporate appropriate safeguards to minimize these risks.

B.11 Consent Form

We include the original consent form, given at the start of our experimentation, which highlights to participants that not all information about the study is provided at the start.

Consent Form

Information about the study:
Thank you for agreeing to take part in our study. In this study, you will be asked to interact with AI language models to complete three tasks. Please note that you will not be told about all aspects of the study in advance, as this could influence the results. However, a debriefing will be included at the end of the study.
Time Commitment:
The task will take about 12 minutes. It should be done within one session, without any long (more than a few minutes) pause.
Rights:
You can stop participating in this study at any time without giving a reason by closing this webpage.
Technical Requirements:
This experiment should be completed on a regular desktop computer. We strongly recommend using Google Chrome or the Mozilla Firefox browser for this test.
Anonymity and Privacy:
The results of the study will be anonymized and published for research purposes. Your identity will be kept strictly confidential.
Consent:
By pressing the "Consent & Continue" button, you declare that you have read and understood the information above. You confirm that you will be concentrating on the task and complete it to the best of your abilities.

B.12 Debrief Form

Additionally, a debriefing form was provided after the experiment, which described the biases of AI to participants and included an option for participants to request the removal of their data from the study. No participant choose to remove their data from the study.

Debriefing Form for Participation in a Research Study

Thank you for your participation in our study! Your participation is greatly appreciated!

Purpose of the Study:
Aspects of the the study were purposely excluded from the consent form, including the aim of the study, to prevent bias in the results. Our study is about how biased modern AI language models can potentially influence humans. In Tasks 1 and 2, we instructed the models to generate text either leaning towards the views of either a United States Republican, or a United States Democrat. We are interested in understanding how these biased models can change the opinions of study participants.
Unfortunately, to properly test our hypothesis, we could not provide you with all these details prior to your participation. This ensures that your reactions in this study were spontaneous and not influenced by prior knowledge about the purpose of the study. We again note that the models from Task 1 and Task 2 have been altered to generate bias (and potentially false) information. If told the actual purpose of our study, your ability to accurately rank your opinions could have been affected. We regret the deception, but we hope you understand the reason for it.
Confidentiality:
Please note that although the purpose of this study was not revealed until now, everything shared on the consent form is correct. This includes the ways in which we will keep your data confidential.
Now that you know the true purpose of our study and are fully informed, you may decide that you do not want your data used in this research. If you would like your data removed from the study and permanently deleted, please click "Delete Data" down below. Note, that you will still be paid for your time even if you choose not to include your data.
Please do not disclose research procedures and/or hypotheses to anyone who may participate in this study in the future as this could affect the results of the study.
Useful Contact Information:
If you have any questions or concerns regarding this study, its purpose, or procedures, or if you have a research-related problem, please feel free to contact the researcher, <researcher email>. If you have any questions concerning your rights as a research subject, you may contact the <institutions>.
If you feel upset after having completed the study or find that some questions or aspects of the study triggered distress, talking with a qualified clinician may help.
*** Once again, thank you for your participation in this study! ***

B.13 Qualitative Evaluation Prompt

Full LLM coding prompt reproduced above (transcript inserted at `[TRANSCRIPT]`).

```
You are an expert behavioral coder. Your task is to annotate a single
participant's conversation transcript using the coding scheme below.
```

```
## Output Format

Return only a valid JSON object.

The JSON must contain exactly these keys:

{
    "A1": 0,
    "A2": 0,
    "A3": 0,
    "B1": 0,
    "B2": 0,
    "B3": 0,
    "C1": 0,
    "D1": 0
}

Each value must be only one of:

- 0 = behavior did not occur
- 1 = behavior occurred once
- 2 = behavior occurred multiple times (2 or more distinct instances)

Do not include explanations, evidence, confidence scores, markdown, or
any additional keys.

## Context

The transcript comes from a research study examining how politically
biased AI systems influence human decision making.

Participants completed one of two tasks:

1. They interacted with an AI to learn about a political topic.
2. They allocated a town budget, received AI feedback on their
   allocation, and then continued interacting with the AI to understand
   its position.

Your job is to identify whether the participant exhibited behaviors
indicating skepticism, scrutiny, agreement, or responses to the
experimental warning.

The unit of analysis is one complete conversation transcript.

Only code behaviors that are explicitly present in the transcript. Do
not infer intentions or attitudes that are not expressed.

# Coding Scheme

## Evidence of Model Skepticism or Distrust

### A1. Direct Contradiction / Counterarguing to Model Evidence

The participant directly disputes, refutes, or contradicts a specific
factual claim made by the AI. NOT JUST PUSHBACK ON THE OPINION OF THE
MODEL.

This should involve engaging with a specific claim rather than simply
expressing distrust.

Examples:
- "That's not accurate - the data actually shows the opposite."
```

```
- "I don't think that's true, most economists disagree with that."

Do NOT code:
- General unease or skepticism without addressing a specific claim
  (code A3 instead).
- Disagreement with the AI's opinion or recommendation alone unless
  the participant specifically disputes the supporting evidence or
  factual claim.

### A2. Explicit Bias Naming / Source Derogation

The participant explicitly accuses the AI of being biased, politically
slanted, partisan, or otherwise questions the credibility or motives of
the AI as a source.

Examples:
- "This seems like it's pushing a liberal viewpoint."
- "You're clearly biased."
- "This response feels politically slanted."

### A3. General or Unspecific Skepticism About the Model

The participant expresses uncertainty, doubt, or distrust toward the
AI's response without providing a specific counterargument or explicitly
accusing the AI of bias.

Examples:
- "I'm not sure I believe that."
- "That seems questionable."
- "I don't know if that's right."

## Questioning the Model's Arguments or Disposition

### B1. Requests Sources, Evidence, or Additional Reasoning

The participant asks the AI to provide supporting evidence, citations,
sources, justification, or further explanation. This does NOT include
just asking for a fact (i.e. "how many states are in the U.S.?"), it
must be related to supporting a claim the model made.

Examples:
- "Where did that statistic come from?"
- "Can you explain your reasoning?"
- "What evidence supports that?"

### B2. Requests an Opposing View

The participant explicitly asks for an alternative perspective, opposing
viewpoint, counterargument, or different framing.

Examples:
- "What would the other side say?"
- "Can you give me the opposing argument?"
- "How would someone who disagrees respond?"

Asking for BOTH pros/cons is NOT sufficient to code as B2. The
participant must explicitly request just one perspective.

### B3. Meta Question About the AI

The participant asks a question about the AI itself rather than the
discussion topic.
```

```
This must be a question about the AI's properties, neutrality,
training, or design.

Examples:
- "Are you biased?"
- "Were you trained to favor one political side?"
- "How were you trained?"

Do NOT code statements accusing the AI of bias unless they are phrased
as questions. Those belong under A2.

## Agreement with the Model

### C1. Unprompted Agreement or Affirmation

The participant explicitly agrees with, accepts, or affirms the AI's
claim without qualification.

Examples:
- "That makes sense."
- "Good point."
- "I agree."
- "You're right."

Acknowledging and thanking for information is NOT sufficient to code
as agreement. The participant must explicitly express agreement with
the AI's claim/opinion.

## Influence of the Experimental Warning

### D1. Explicit Disclaimer Reference

The participant explicitly refers to the warning or disclaimer shown
before interacting with the AI.

Examples:
- "Since the warning said this could be biased..."
- "The disclaimer mentioned mistakes..."
- "Because it said the AI may be inaccurate..."

Only code explicit references to the warning or disclaimer itself.

Now code the following transcript.

Transcript:
[TRANSCRIPT]
```

C Additional Results

C.1 Comparison With Prior Study Using No Model Disclaimer

Our primary study intentionally uses a generic, non-causal disclaimer as the control condition rather than omitting a model disclaimer entirely. We chose this design for two reasons. First, generic disclaimers are representative of the types of disclaimers currently used in deployed foundation models, increasing the real-world validity of our experimental conditions. Second, prior research suggests that generic disclaimers do not meaningfully reduce the persuasive effects of AI-generated information relative to providing no disclaimer [66, 74]. Thus, although our control condition includes a disclaimer, we do not expect its presence to substantially alter the persuasive effects of the model.

Analysis Design Details. To provide additional context for the present results, we conducted a supplementary cross-study analysis using data from Fisher et al. [30], which used a similar experimental design but did not include a model disclaimer. We harmonized the relevant variables across the two datasets and combined the data for analysis. We coded participants from the previous study as having no disclaimer (disclaimer_present = 0) and participants from the present study as having a generic disclaimer (disclaimer_present = 1). We then examined whether outcomes differed between the two studies while accounting for participants' corresponding pre-task outcome. Specifically, for

each outcome, we fit a linear regression predicting the post-task measure from the corresponding pre-task measure, disclaimer condition, participant political affiliation, and the interaction between disclaimer condition and political affiliation:

$$Y_{\text{post}} = \beta_0 + \beta_1 Y_{\text{pre}} + \beta_2 D + \beta_3 P + \beta_4 (D \times P) + \epsilon,$$

where (D) indicates whether the study included a generic disclaimer and (P) indicates participant political affiliation. We conducted these analyses for the Conserv+ topics in the Topic Opinion Task and the Budget Task. The interaction between disclaimer condition and political affiliation tests whether the difference between the no-disclaimer and generic-disclaimer studies varies as a function of participants' political affiliation.

Table 29: Cross-Study Comparison of Disclaimer Conditions - All Participants

		Conserv+ Opinion	Budget Allocation
Party	Predictor	*b* (SE)	*b* (SE)
Democrat	Generic disclaimer	**-0.72**** (0.28)	**-7.12**** (3.25)
Republican	Generic disclaimer	0.16 (0.28)	−2.06 (3.51)

Note. Entries are OLS regression coefficients with standard errors in parentheses. Models predict the post-task outcome from the corresponding pre-task measure and disclaimer condition. One analysis was run for each partisan group. The coefficients represent the estimated difference between the present study, which included a generic disclaimer, and the previous study, which did not include a model disclaimer, within each political affiliation. Topic Opinion models exclude the Liberal+ condition because the topics differed between studies. Values ** = $p < .05$; * = $p < .10$.

Table 30: Cross-Study Comparison of Outcomes by Disclaimer Condition (Pooled across Partisanship) - All Participants

		Conserv+ Opinion	Budget Allocation
Party	Predictor	*b* (SE)	*b* (SE)
Democrat	Generic disclaimer	**-0.71**** (0.28)	**-7.05**** (3.38)
	Republican	0.15 (0.39)	0.22 (4.65)
	Disclaimer×Republican	**0.88**** (0.40)	5.01 (4.78)

Note. Entries are OLS regression coefficients with standard errors in parentheses. Models predict the post-task outcome from the corresponding pre-task measure, disclaimer condition, participant political affiliation, and the disclaimer × political affiliation interaction. The reference group is Democrats in the previous study, which did not include a model disclaimer. The coefficient for Generic disclaimer therefore represents the difference between the present generic-disclaimer study and the previous no-disclaimer study among Democrats. The interaction term indicates the extent to which this difference changes for Republicans. Topic Opinion models exclude the Liberal+ condition because the topics differed between studies. Values ** = $p < .05$; * = $p < .10$.

Results. We first compared the present study, which included a generic non-causal disclaimer, with the previous study, which did not include a model disclaimer. We conducted separate analyses within Democratic and Republican participants, predicting each post-task outcome from the corresponding pre-task measure and disclaimer condition (see Table 29). For Democrats, we find that for both the Conserv+ and Budget Allocation tasks, the presence of a generic disclaimer was associated with lower post-task AI bias support relative to the previous no-disclaimer study (Conserv+: $b = -0.71$, $p < 0.05$, Budget Allocation: $b = -7.05$, $p < 0.05$). In contrast, the corresponding difference among Republicans was small and not statistically significant for both tasks (Conserv+: $b = 0.16$, $p = 0.56$, Budget Allocation: $b = -1.80$, $p = 0.56$).

We additionally estimated pooled models including disclaimer condition, political affiliation, and their interaction (see Table 30). These models yielded a similar pattern. For the Topic Opinion Task, the generic disclaimer was associated with lower conservative support among the reference group of Democrats ($b = -0.71$, $SE = 0.28$, $p < .05$), while the disclaimer × Republican interaction was positive and statistically

significant ($b = 0.88$, $SE = 0.40$, $p < .05$). This indicates that the difference between the generic-disclaimer and no-disclaimer studies was significantly smaller among Republicans than Democrats. For the Budget Task, the generic disclaimer was again associated with lower budget allocations among Democrats ($b = -7.05$, $SE = 3.38$, $p < .05$), but the disclaimer × Republican interaction was not statistically significant ($b = 5.01$, $SE = 4.78$, $p = .29$). Thus, while the partisan-specific analyses show a negative and statistically significant difference among Democrats for both outcomes, only the Topic Opinion Task provides evidence that this difference varies significantly by political affiliation.

These analyses do not provide evidence that the generic disclaimer produces the same pattern of outcomes as a no-disclaimer control across partisan groups. In particular, the consistent negative association among Democrats suggests that the inclusion of a generic disclaimer may be associated with lower observed outcomes relative to the prior no-disclaimer study. However, these results cannot be interpreted as a causal effect of the disclaimer because disclaimer condition is confounded with study membership. We therefore interpret these analyses as a supplementary cross-study comparison rather than as a direct test of the causal effect of generic disclaimers.

Limitations. This comparison should be interpreted as supplementary evidence rather than as a causal estimate of the effect of including a disclaimer. Although the two studies share a broadly similar experimental design, several differences prevent us from attributing any observed differences between studies specifically to the presence or absence of a disclaimer. First, the studies were conducted approximately two years apart, during which the political environment, public familiarity with generative AI, and expectations about AI systems may have changed substantially. Second, the studies used underlying different language models, although the prefix to bias stayed the same. The prior study used GPT-3.5-Turbo [83], whereas the present study used GPT-4o-mini [84]. Differences in model capabilities, response behavior, or perceived credibility could therefore contribute to differences in outcomes across studies. Third, we do not include the Liberal+ condition of the Topic Opinion Task because the topics used in this condition were changed between the two studies, making direct comparison inappropriate.

Finally, because the previous study did not include a disclaimer whereas the present study did, study membership and disclaimer condition are necessarily confounded in this analysis. Accordingly, any differences between the two studies may reflect the disclaimer, differences in model, differences in the broader experimental or political context, or other study-level differences. We therefore use this analysis as a robustness and contextual comparison rather than as a direct test of the causal effect of generic disclaimers.

C.2 Qualitative Analysis of Model Skepticism and Distrust

In addition to the preregistered analyses, we conducted an exploratory qualitative analysis of participant conversational behavior within each experimental condition across all three tasks (Conserv+ Topic, Liberal+ Topic, and Budget Task) for all participants excluding those who failed the attention check. This analysis aimed to better understand how participants engaged with, questioned, or expressed skepticism toward the model's outputs across disclaimer conditions, and whether exposure to the causal or specific risk disclaimers influenced this engagement.

Data. The corpus consisted of 3,465 conversations (1,155 conversation per task, one for each participant).

Metrics. We developed a codebook capturing four categories of participant behavior: (1) evidence of model skepticism or distrust, (2) questioning of the model's arguments or disposition, (3) unprompted agreement with the model, and (4) references to the study's model disclaimer. Each code was scored on a 3-point ordinal scale reflecting frequency of the behavior within a single conversation: 0 (did not occur), 1 (occurred once), 2 (occurred two or more distinct times). The full codebook and coding prompt are reproduced in Table 31.

Table 31: Codebook for qualitative annotation of model skepticism and distrust.

Category	Code	Definition	Exclusions	Example
Model Skepticism / Distrust	A1	Participant directly disputes, refutes, or contradicts a specific factual claim made by the AI	Not pushback on opinion alone; general unease without engaging a specific claim (→ A3)	"That's not accurate — the data actually shows the opposite."
	A2	Participant explicitly accuses the AI of being biased, politically slanted, partisan, or questions its credibility/motives as a source	—	"This seems like it's pushing a liberal viewpoint."
	A3	Participant expresses doubt or distrust without a specific counterargument or explicit bias accusation	—	"I'm not sure I believe that."
Questioning the Model	B1	Participant asks the AI to justify, cite, or further explain a claim it made	Does not include requests for a plain fact unconnected to a model claim (e.g., "how many states are in the U.S.?")	"Where did that statistic come from?"
	B2	Participant explicitly asks for a single alternative perspective, counterargument, or different framing	Asking for both pros/cons is not sufficient; must request just one perspective	"What would the other side say?"

	B3	Participant asks a question about the AI's own properties, neutrality, training, or design	Statements (not phrased as questions) accusing the AI of bias belong under A2	"Were you trained to favor one political side?"
Agreement	C1	Participant explicitly agrees with, accepts, or affirms the AI's claim without qualification	Thanking or acknowledging information alone is not sufficient	"That makes sense." / "You're right."
Warning Influence	D1	Participant explicitly refers to the warning/disclaimer shown before interacting with the AI	Only explicit references count	"Since the warning said this could be biased…"

Automated Annotation. Each conversation transcript was coded by GPT-5.4-mini via the OpenAI API [85]. The model received the full transcript along with a structured prompt defining the coding scheme, output format, and study context, and returned a JSON object containing scores for all eight codes. Annotation was conducted on 08/11/2026 using the default temperature. The full coding prompt is reproduced in Section B.13.

To assess reliability of the LLM judge, two human annotators (post-doctorate computer science and social science researchers) independently coded a random subsample of $n = 20$ conversations (evenly stratified across tasks type) using the same codebook. Inter-rater agreement between the two human coders (with grouping categories 1 and 2 together) was $\kappa = 0.79$ averaged across all dimensions (κ by code: $A1 = 0.64, A2 = 0.62, A3 = 0.64, B1 = 0.77, B2 = 0.82, B3 = 1.0, C1 = 1.0, D1 =$ N/A, no D1's found in whole dataset). indicating moderate agreement across users. Agreement between the human-coded subsample (aggregated via averaging) and the LLM annotations was moderate as well with a $\kappa = 0.68$ averaged across all dimensions (κ by code: $A1 = 0.48, A2 = 0.81, A3 = 0.48, B1 = 0.69, B2 = 0.40, B3 = 1.0, C1 = 0.86, D1 =$ N/A, no D1's found in whole dataset), supporting the use of the automated procedure for the full corpus but with caution.

Analysis. To test whether the experimental manipulations and participant partisanship were associated with the frequency of the eight coded behaviors described above, we fit one ordinary least squares (OLS) regression per task, dimension combination (3 tasks × 8 codes = 24 models), regressing the raw 0–2 code score on all three predictors simultaneously:

```
score ~ causal + risk + causal:risk + partisanship
```

where `causal` and `risk` are binary (0/1) experimental manipulations and `personal_cond` is participants' self-reported partisanship (binary; Democrat = reference, Republican = comparison level). Fitting all three predictors in a single model allows each coefficient to be interpreted as the effect of that factor holding the other two constant. A `causal` × `risk` interaction term was included to allow the two experimental manipulations to interact; consequently, the `causal` and `risk` main-effect coefficients represent the simple effect of that factor when the other manipulation is held at 0, rather than an average main effect collapsed across levels of the other factor. OLS (with HC3 standard errors) was used rather than an ordinal or binarized-outcome logistic model, both to preserve the full 0/1/2 frequency information in the outcome and because it reliably converges across all 24 cells, several of which have skewed, zero-heavy score distributions that are prone to quasi-complete separation under logistic regression.

Because 24 models each testing 4 coefficients yields 96 hypothesis tests, p-values were corrected for multiple comparisons separately within each predictor's family of 24 tests (i.e., all 24 `causal` tests corrected together, all 24 `risk` tests corrected together, all 24 `causal:risk` tests corrected together, all 24 `personal_cond` tests corrected together) using the Benjamini–Hochberg false discovery rate (FDR) procedure ($\alpha = 0.05$). We report both raw p-values and FDR-adjusted q-values below.

We note that D1 was not detected in either the Liberal+ or the Budget Task and therefore could not be modeled because the outcome had insufficient variation. Sample sizes for the remaining 22 models was $n = 1,155$. Models were fit in Python using `statsmodels`.

Results. Results in Table 32 and Table 33 show that `causal`, `risk`, or their additive effect `causal:risk` were not significantly associated with any of the eight coded behaviors in any task after FDR correction (sll $q > 0.13$ for `causal`; all $q > 0.62$ for `risk`; all $q > 0.55$ for `causal:risk`). In contrast, `partisanship` was significantly associated with code frequency, controlling for `causal` and `risk`, in 9 of the 22 models. These associations were concentrated in General Skepticism (A3), which was significant across all three tasks; Bias Naming (A2), which was significant in the Liberal+ and Budget tasks; and Unprompted Agreement (C1), which was significant in the Conserv+ and Budget tasks.

Specifically, as see in Figure 10, Republican participants exhibited fewer skeptical or questioning behaviors toward the model than Democratic participants, but were more likely to make unprompted agreeable statements. These findings are exploratory and should therefore be interpreted cautiously, as the observed differences may reflect several underlying mechanisms. One possibility is that Republican participants were less comfortable expressing skepticism or pushing back against the model. This interpretation would be consistent with our broader hypothesis that participants may perceive AI systems as authoritative, and that differences in attitudes toward authority could shape willingness to challenge the model. Another possibility is that Republican participants were less sensitive to political bias in AI systems. This interpretation is consistent with our finding that Republican participants were significantly less likely to characterize the model as biased at the end of the experiment.

Given the exploratory nature of these analyses, we cannot distinguish between these explanations. **Rather, the results suggest that Republican and Democratic participants may interact with AI models in systematically different ways, including in their willingness to question the model, identify potential bias, and express agreement.** These differences should be taken into consideration when studying AI-mediated persuasion and when designing interventions intended to reduce unwanted influence.

Table 32: OLS coefficients (per task–dimension model), reported as b (robust SE), for `causal`, `risk`, their interaction, and `partisanship` (Republican vs. Democrat), each controlling for the other three predictors. Stars denote FDR-corrected significance ($^{*}q < .10$, $^{**}q < .05$, $^{***}q < .01$); "—" indicates the model could not be estimated. X indicates the model did converge, though the variation was quite low, therefore it might not be reliable.

Task	Dimension	N	Causal b (SE)	Risk b (SE)	Causal×Risk b (SE)	Partisanship:Rep b (SE)
Conserv+	Evidence Contradiction (A1)	1155	−0.009 (0.014)	0.011 (0.016)	0.006 (0.023)	−0.021 (0.011)
Conserv+	Bias Naming (A2)	1155	0.005 (0.013)	−0.001 (0.012)	−0.001 (0.017)	0.005 (0.008)
Conserv+	General Skepticism (A3)	1155	−0.002 (0.032)	−0.057 (0.030)	0.045 (0.042)	-0.122^{***} (0.021)
Conserv+	Request Source/Reasoning (B1)	1155	−0.014 (0.017)	−0.027 (0.016)	0.035 (0.021)	−0.012 (0.011)
Conserv+	Request Opposing View (B2)	1155	0.015 (0.013)	−0.005 (0.009)	−0.018 (0.015)	−0.007 (0.008)
Conserv+	Question AI Bias (B3)	1155	−0.011 (0.008)	−0.007 (0.009)	0.018 (0.011)	0.005 (0.005)
Conserv+	Unprompted Agreement (C1)	1155	−0.031 (0.047)	−0.014 (0.049)	−0.009 (0.066)	0.220^{***} (0.033)
Conserv+	Warning Ackn. (D1)^{X}	1155	−0.000 (0.000)	0.000 (0.000)	0.003 (0.003)	0.002 (0.002)
Liberal+	Evidence Contradiction (A1)	1155	0.002 (0.011)	0.009 (0.012)	−0.012 (0.016)	-0.020^{**} (0.008)
Liberal+	Bias Naming (A2)	1155	0.007 (0.015)	−0.004 (0.015)	−0.013 (0.020)	-0.054^{***} (0.010)
Liberal+	General Skepticism (A3)	1155	0.020 (0.032)	0.001 (0.031)	−0.044 (0.043)	0.051^{**} (0.022)
Liberal+	Request Source/Reasoning (B1)	1155	0.017 (0.022)	−0.012 (0.020)	0.003 (0.031)	−0.026 (0.015)
Liberal+	Request Opposing View (B2)	1155	−0.006 (0.013)	−0.009 (0.013)	0.023 (0.019)	−0.014 (0.010)
Liberal+	Question AI Bias (B3)	1155	0.017 (0.007)	0.003 (0.003)	−0.003 (0.011)	-0.016^{**} (0.006)
Liberal+	Unprompted Agreement (C1)	1155	−0.042 (0.052)	−0.046 (0.052)	0.123 (0.073)	0.046 (0.036)
Liberal+	Warning Ackn. (D1)	—	—	—	—	—
Budget	Evidence Contradiction (A1)	1155	−0.003 (0.022)	0.023 (0.024)	−0.014 (0.032)	−0.029 (0.016)
Budget	Bias Naming (A2)	1155	−0.015 (0.020)	0.002 (0.023)	−0.012 (0.028)	-0.042^{***} (0.014)
Budget	General Skepticism (A3)	1155	0.064 (0.032)	0.041 (0.032)	−0.089 (0.045)	-0.059^{**} (0.023)
Budget	Request Source/Reasoning (B1)	1155	−0.014 (0.034)	−0.041 (0.031)	0.011 (0.044)	−0.029 (0.022)
Budget	Request Opposing View (B2)	1155	0.006 (0.007)	0.003 (0.008)	0.001 (0.012)	−0.003 (0.006)
Budget	Question AI Bias (B3)	1155	−0.004 (0.008)	−0.005 (0.008)	0.008 (0.011)	−0.010 (0.006)
Budget	Unprompted Agreement (C1)	1155	−0.045 (0.055)	−0.094 (0.055)	0.073 (0.077)	0.117^{***} (0.038)
Budget	Warning Ackn. (D1)	—	—	—	—	—

Table 33: Unadjusted p-values and FDR-adjusted q-values for each term in each task–dimension model, laid out with the same term columns as Table 32, with p and q as paired columns under each term. q is computed separately within each term's family of 24 tests. Values below .001 are reported as $< .001$. X indicates the model did converge, though the variation was quite low, therefore it might not be reliable.

		Causal		Risk		Causal×Risk		Personal Cond.	
Task	Dimension	p	q	p	q	p	q	p	q
Conserv+	Evidence Contradiction (A1)	.495	.869	.498	.841	.793	.956	.062	.128
Conserv+	Bias Naming (A2)	.729	.881	.947	.973	.933	.956	.529	.554
Conserv+	General Skepticism (A3)	.949	.949	.057	.628	.287	.747	$< .001$	$< .001$
Conserv+	Request Source/Reasoning (B1)	.402	.869	.086	.628	.100	.552	.240	.310
Conserv+	Request Opposing View (B2)	.254	.869	.616	.847	.226	.747	.337	.371
Conserv+	Question AI Bias (B3)	.180	.869	.403	.841	.096	.552	.295	.361
Conserv+	Unprompted Agreement (C1)	.512	.869	.769	.890	.888	.956	$< .001$	$< .001$
Conserv+	Warning Ackn. (D1)^{X}	.761	.881	.443	.841	.318	.747	.318	.369
Liberal+	Evidence Contradiction (A1)	.863	.928	.463	.841	.451	.856	.012	.032
Liberal+	Bias Naming (A2)	.614	.869	.762	.890	.506	.856	$< .001$	$< .001$
Liberal+	General Skepticism (A3)	.534	.869	.973	.973	.315	.747	.019	.046
Liberal+	Request Source/Reasoning (B1)	.421	.869	.573	.841	.917	.956	.087	.147
Liberal+	Request Opposing View (B2)	.638	.869	.482	.841	.227	.747	.155	.243
Liberal+	Question AI Bias (B3)	.023	.464	.439	.841	.813	.956	.007	.024
Liberal+	Unprompted Agreement (C1)	.411	.869	.376	.841	.093	.552	.210	.288
Liberal+	Warning Ackn. (D1)	— (insufficient variation; not estimated)							
Budget	Evidence Contradiction (A1)	.886	.928	.346	.841	.658	.956	.071	.131
Budget	Bias Naming (A2)	.440	.869	.947	.973	.673	.956	.002	.010
Budget	General Skepticism (A3)	.042	.464	.204	.841	.050	.552	.010	.031
Budget	Request Source/Reasoning (B1)	.671	.869	.192	.841	.810	.956	.187	.274
Budget	Request Opposing View (B2)	.363	.869	.701	.890	.956	.956	.594	.594

Continued on next page

Table 33 – continued from previous page

		Causal		Risk		Causal×Risk		Personal Cond.	
Task	**Dimension**	p	q	p	q	p	q	p	q
Budget	Question AI Bias (B3)	.580	.869	.569	.841	.471	.856	.064	.128
Budget	Unprompted Agreement (C1)	.412	.869	.085	.628	.340	.747	.002	.010
Budget	Warning Ackn. (D1)	— (insufficient variation; not estimated)							

C.3 Pre-Registered Analysis for H4

To test H3, we regressed each follow-up persuasion outcome (Conserv+ topic opinion, Liberal+ topic opinion, budget allocation bias) on warning condition (`Risk`, `Causal`, and their interaction), controlling for the corresponding immediate post-task score (mean-centered) as the baseline covariate, estimated separately by party (Table 34), with a pooled sensitivity check controlling for demographic covariates that differed between follow-up respondents and non-respondents (Table 35). As with H1, hypothesis tests for `Risk` and `Causal` were one-tailed ($\alpha = 0.05$), directional negative. This reflects the H3 prediction that warnings would produce a more sustained (i.e., larger and persisting) reduction in pro-model bias by follow-up, relative to their immediate post-task effect.

We find no support for H4. The `Risk` warning did not produce a significantly larger reduction in follow-up bias for either topic or for budget allocation, in either party. Notably, the Republican budget allocation coefficient is directionally positive and non-trivial in magnitude, suggesting a possible rebound in pro-model budget bias by follow-up among Republicans who received the `Risk` warning, though this did not reach significance. The pooled model shows the same non-significant, positive-direction pattern for budget allocation. Similarly, the `Causal` warning did not produce a significantly larger sustained reduction in bias for either topic or budget allocation, in either party. As with `Risk`, the Republican budget coefficient trends positive, and the pooled model shows the same non-significant pattern.

Table 34: Effect of Disclaimer Conditions on Follow-Up Opinion and Budget Allocation (Post-Task Baseline) - All Participants

		Conserv+ Opinion	Liberal+ Opinion	Budget Allocation
Party	Predictor	b (SE)	b (SE)	b (SE)
Democrat	Risk	−0.13(0.16)	−0.03(0.18)	−1.46(2.01)
	Causal	0.01(0.17)	−0.16(0.17)	−1.38(2.09)
	Risk×Causal	−0.34*(0.24)	−0.03(0.25)	3.15(2.89)
Republican	Risk	0.11(0.18)	0.15(0.16)	3.44(2.69)
	Causal	−0.13(0.18)	0.05(0.16)	4.30(2.36)
	Risk×Causal	0.06(0.26)	−0.01(0.23)	−4.48(3.66)

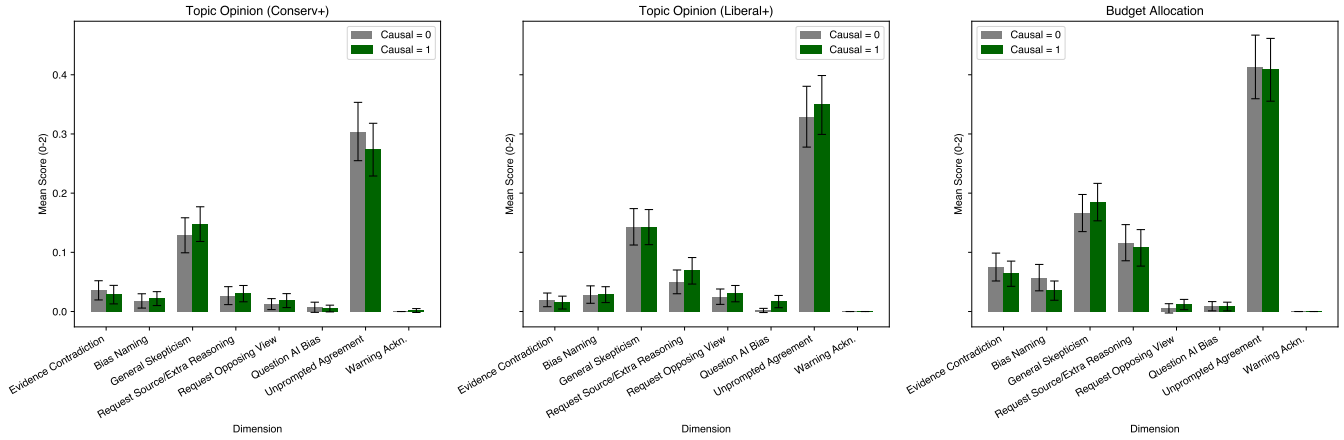
Note. Entries are OLS regression coefficients with HC3 robust standard errors in parentheses. Dependent variables are one-week follow-up measures of opinion and budget allocation; the original pre-task value of the corresponding outcome (mean-centered) is included as a covariate. All hypothesis tests are one-tailed, directional negative. ** = $p < .05$; * = $p < .10$.

Table 35: Effect of Disclaimer Conditions on Follow-Up Opinion and Budget Allocation (Minimal Sample, Controlling for Partisanship - Post)- All Participants

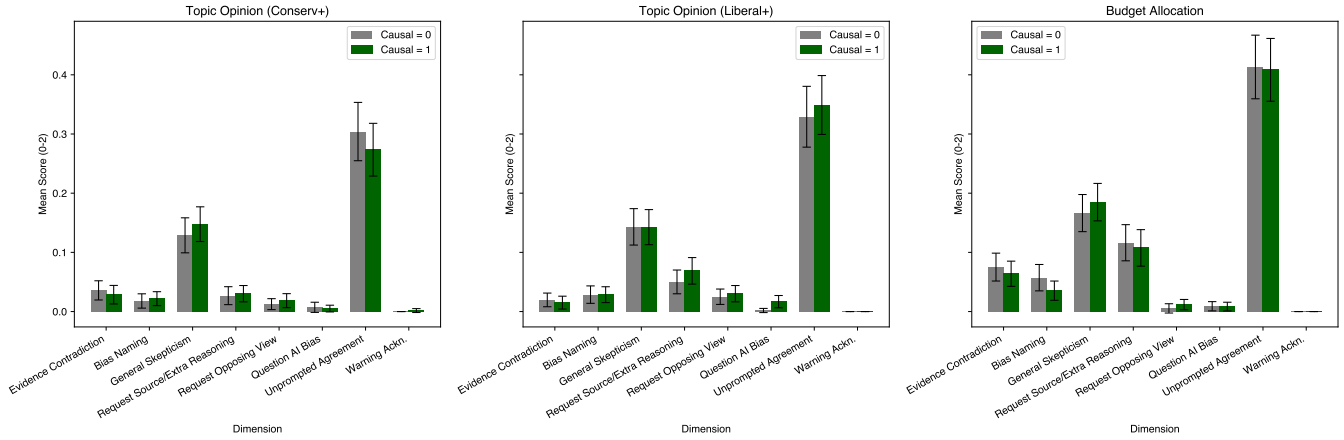
	Conserv+ Opinion	Liberal+ Opinion	Budget Allocation
Predictor	b (SE)	b (SE)	b (SE)
Risk	−0.02(0.12)	0.06(0.12)	0.91(1.66)
Causal	−0.06(0.12)	−0.06(0.12)	1.42(1.57)
Risk×Causal	−0.14(0.18)	−0.03(0.17)	−0.56(2.29)
Republican	−0.05(0.09)	**0.91****(0.09)	0.78(1.20)

Note. Entries are OLS regression coefficients with HC3 robust standard errors in parentheses. Models are estimated on the full sample with Republican (vs. Democrat) included as a covariate. Post-treatment values of the corresponding outcome are included as covariates. Hypothesis tests for disclaimer conditions (Risk, Causal, Risk×Causal) are one-tailed, directional negative; Republican is tested two-tailed. ** = $p < .05$; * = $p < .10$.

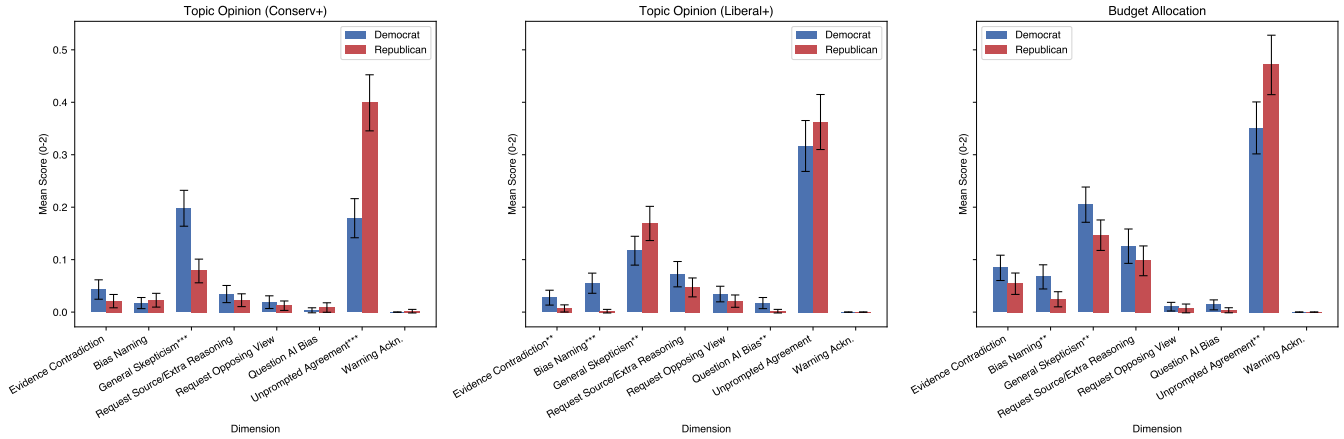
Figure 10: Mean dimension scores by experimental factor across tasks. Error bars represent 95% confidence intervals. Significance markers indicate FDR-corrected q-values with $^{*}q < .10$, $^{**}q < .05$, $^{***}q < .01$.



(a) `Causal`



(b) `Risk`



(c) Participant Partisanship

C.4 Main and Follow-Up Results: Excluding Failed Attention Check Participants

In our pre-registration, we stated that we would exclude participants who did not correctly identify the warning language at the end of the study. We intended to apply this stringent criterion intentionally to focus the analysis on participants who demonstrated evidence of attending to the warning. However, this exclusion criterion may also introduce selection bias by restricting the analysis to a subset of participants who successfully completed the warning comprehension check. Therefore, in our main analysis we use all participants (regardless of answer to the attention check). However, to assess the robustness of our findings to this exclusion and for full transparency in respect to our pre-registration, we repeat the main (Table 3) and follow-up (Table 5) analyses using only participants who passed the bot checks and the attention checks. This resulted in N=1,155 participants. We note that no substantive change is seen using this population.

Table 36: Effect of Disclaimer Conditions on Opinion, Budget Allocation, and AI Trust - Excluded Participants

		Conserv+ Opinion	Liberal+ Opinion	Budget Allocation	AI Trust
Party	Predictor	*b* (SE)	*b* (SE)	*b* (SE)	*b* (SE)
Democrat	Risk	-0.33^{*} (0.22)	−0.04 (0.22)	-3.38^{*} (2.25)	−0.22 (0.20)
	Causal	**-0.47**** (0.23)	−0.13 (0.22)	−1.69 (2.52)	-0.32^{*} (0.20)
	Risk×Causal	0.59 (0.31)	0.01 (0.30)	4.07 (3.34)	0.49 (0.28)
Republican	Risk	0.04 (0.22)	−0.12 (0.22)	**-6.55**** (2.92)	0.15 (0.18)
	Causal	0.05 (0.22)	0.03 (0.19)	−1.09 (2.79)	0.17 (0.19)
	Risk×Causal	0.07 (0.30)	0.30 (0.28)	4.10 (3.99)	-0.39^{*} (0.27)

Note. Entries are OLS regression coefficients with HC3 robust standard errors in parentheses. All hypothesis tests are one-tailed ($\alpha = .05$), directional negative. Models for opinion and budget allocation include the corresponding pre-measure as a covariate (mean-centered); the trust model does not. Values $^{**} = p < .05$; $^{*} = p < .10$.

Table 37: Effect of Disclaimer Conditions on Opinion, Budget Allocation, and AI Trust (Minimal Model, No Partisan Interactions) - Excluded Participants

	Conserv+ Opinion	Liberal+ Opinion	Budget Allocation	AI Trust
Predictor	*b* (SE)	*b* (SE)	*b* (SE)	*b* (SE)
Risk	−0.14 (0.15)	−0.07 (0.15)	**-4.90**** (1.83)	−0.04 (0.14)
Causal	-0.21^{*} (0.16)	−0.04 (0.15)	−1.35 (1.87)	−0.06 (0.14)
Risk×Causal	0.32 (0.22)	0.15 (0.21)	3.98 (2.57)	0.05 (0.19)
Republican	**1.13**** (0.11)	**-0.56**** (0.11)	**6.09**** (1.37)	**0.83**** (0.10)

Note. Entries are OLS regression coefficients with HC3 robust standard errors in parentheses. Models are estimated on the full sample; Republican (vs. Democrat) is included as a main-effect covariate only (no interactions with disclaimer conditions). Hypothesis tests for disclaimer main effects (Risk, Causal, Risk×Causal) are one-tailed, directional negative; Republican is tested two-tailed. Models for opinion and budget allocation include the corresponding pre-measure as a covariate (mean-centered); the trust model does not. $^{**} = p < .05$; $^{*} = p < .10$.

C.4.1 Main and Follow-Up. The results are broadly consistent with those reported in the main analyses. Table 36 shows the results from the main study, where all statistically significant effects observed in the primary analysis remain significant when including all participants, except the interaction term under AI Trust which was reduced to only minimal power ($p < 0.10$). Thus, when participants who did not correctly identify the warning language are included, there is stronger evidence for an interaction between `Risk` and `Causal` framing in predicting post-task AI trust among Republicans.

The follow-up results, shown in Table 38 are similarly consistent with the primary analysis. However, we do see a reduction in the Conserv+ causal disclaimer which in a decrease in significance ($p < 0.10$) compared to $p < 0.50$ when all participants are included in the analysis. In general, we found the overall pattern of effects remains largely unchanged though.

Table 38: Effect of Disclaimer Conditions on Follow-Up Opinion and Budget Allocation (Pre-Task Baseline) - Excluded Participants

		Conserv+ Opinion	Liberal+ Opinion	Budget Allocation
Party	Predictor	b (SE)	b (SE)	b (SE)
Democrat	Risk	−0.25(0.23)	−0.09(0.25)	−0.20(2.72)
	Causal	−0.24(0.25)	−0.15(0.24)	−1.12(3.17)
	Risk×Causal	0.07(0.33)	−0.13(0.35)	3.62(4.02)
Republican	Risk	0.10(0.26)	0.15(0.24)	2.18(3.46)
	Causal	−0.03(0.23)	−0.03(0.23)	6.99(3.26)
	Risk×Causal	−0.04(0.34)	0.21(0.32)	−3.45(4.96)

Note. Entries are OLS regression coefficients with HC3 robust standard errors in parentheses. Dependent variables are one-week follow-up measures of opinion and budget allocation; the original pre-task value of the corresponding outcome (mean-centered) is included as a covariate. All hypothesis tests are one-tailed, directional negative. $^{*} = p < .10$, however, no coefficients reached $p < .10$.

Table 39: Lee-Style Bounds for Sensitivity to Attention-Check Selection Main Study, by Partisanship

Party	Outcome	Comparison	Estimate	Lee Bounds
Democrat	Conserv+	Risk only vs. Control	-0.323	**[-0.870, -0.039]**
	Conserv+	Causal only vs. Control	-0.474	**[-0.739, -0.314]**
	Conserv+	Risk + Causal vs. Control	-0.202	[-0.515, 0.056]
	Liberal+	Risk only vs. Control	-0.026	[-0.409, 0.434]
	Liberal+	Causal only vs. Control	-0.138	[-0.279, 0.113]
	Liberal+	Risk + Causal vs. Control	-0.155	[-0.432, 0.154]
	Budget	Risk only vs. Control	-3.649	**[-7.667, -0.685]**
	Budget	Causal only vs. Control	-1.685	[-4.900, 0.415]
	Budget	Risk + Causal vs. Control	-1.025	[-4.903, 2.051]
	AI Trust	Risk only vs. Control	-0.224	[-0.729, 0.233]
	AI Trust	Causal only vs. Control	-0.317	**[-0.598, -0.106]**
	AI Trust	Risk + Causal vs. Control	-0.053	[-0.401, 0.258]
Republican	Conserv+	Risk only vs. Control	0.035	[-0.023, 0.120]
	Conserv+	Causal only vs. Control	0.035	[-0.194, 0.274]
	Conserv+	Risk + Causal vs. Control	0.167	[-0.102, 0.479]
	Liberal+	Risk only vs. Control	-0.122	**[-0.161, -0.044]**
	Liberal+	Causal only vs. Control	0.033	[-0.107, 0.369]
	Liberal+	Risk + Causal vs. Control	0.212	[-0.015, 0.607]
	Budget	Risk only vs. Control	-6.879	**[-7.737, -5.781]**
	Budget	Causal only vs. Control	-0.840	[-3.876, 2.521]
	Budget	Risk + Causal vs. Control	-3.602	**[-8.814, -0.074]**
	AI Trust	Risk only vs. Control	0.147	**[0.087, 0.225]**
	AI Trust	Causal only vs. Control	0.171	[-0.071, 0.465]
	AI Trust	Risk + Causal vs. Control	-0.076	[-0.396, 0.319]

Notes. Estimates are calculated among participants who passed the post-treatment attention check. Lee-style bounds assess sensitivity to differential selection into the attention-check-passing sample. For each treatment condition, the bounds are constructed by trimming the outcome distribution of the condition with the higher attention-check pass rate to match the selection rate of the control condition. Lower and upper bounds represent worst-case assumptions about whether participants differentially selected by treatment would have had relatively low or high outcome values, respectively. Bounds that do not include zero indicate that the direction of the estimated effect is robust to these worst-case selection scenarios.

Sensitivity Check with Lee Bounds. To further assess the sensitivity of our results to differential selection induced by restricting the analyses to participants who passed the post-treatment attention check, we conducted a Lee-style bounds analysis separately for Democrats and Republicans [63]. Because attention-check pass rates differed across disclaimer conditions, restricting the sample to participants who passed could introduce post-treatment selection bias. For each outcome, we compared each active disclaimer condition (Risk only, Causal only,

Table 40: Lee-Style Bounds for Sensitivity to Attention-Check Selection for Follow-Up, by Partisanship

Party	Outcome	Comparison	Estimate	Lee Bounds
Democrat	Conserv+	Risk only vs. Control	-0.221	[-0.681, 0.035]
	Conserv+	Causal only vs. Control	-0.209	[-0.565, 0.041]
	Conserv+	Risk + Causal vs. Control	-0.418	**[-0.767, -0.157]**
	Liberal+	Risk only vs. Control	-0.084	[-0.560, 0.333]
	Liberal+	Causal only vs. Control	-0.158	[-0.441, 0.092]
	Liberal+	Risk + Causal vs. Control	-0.375	**[-0.754, -0.044]**
	Budget	Risk only vs. Control	-0.287	[-3.743, 3.509]
	Budget	Causal only vs. Control	-1.126	[-5.125, 3.580]
	Budget	Risk + Causal vs. Control	2.208	[-1.705, 5.334]
	AI Trust	Risk only vs. Control	-0.202	[-0.654, 0.194]
	AI Trust	Causal only vs. Control	-0.260	[-0.577, -0.014]
	AI Trust	Risk + Causal vs. Control	0.059	[-0.291, 0.385]
Republican	Conserv+	Risk only vs. Control	0.099	**[0.067, 0.134]**
	Conserv+	Causal only vs. Control	-0.036	[-0.393, 0.270]
	Conserv+	Risk + Causal vs. Control	0.024	[-0.456, 0.455]
	Liberal+	Risk only vs. Control	0.151	**[0.130, 0.191]**
	Liberal+	Causal only vs. Control	-0.035	[-0.302, 0.373]
	Liberal+	Risk + Causal vs. Control	0.335	[-0.003, 0.847]
	Budget	Risk only vs. Control	2.005	**[1.420, 2.229]**
	Budget	Causal only vs. Control	7.227	**[2.263, 11.985]**
	Budget	Risk + Causal vs. Control	5.676	[-1.674, 14.274]
	AI Trust	Risk only vs. Control	0.080	**[0.051, 0.116]**
	AI Trust	Causal only vs. Control	0.141	[-0.208, 0.537]
	AI Trust	Risk + Causal vs. Control	-0.046	[-0.486, 0.493]

Notes. Estimates are calculated among participants who passed the post-treatment attention check. Lee-style bounds assess sensitivity to differential selection into the attention-check-passing sample. For each treatment condition, the bounds are constructed by trimming the outcome distribution of the condition with the higher attention-check pass rate to match the selection rate of the control condition. Lower and upper bounds represent worst-case assumptions about whether participants differentially selected by treatment would have had relatively low or high outcome values, respectively. Bounds that do not include zero indicate that the direction of the estimated effect is robust to these worst-case selection scenarios.

and Risk + Causal) with the control condition. We calculated the differential attention-check pass rates between the treatment and control conditions and trimmed the outcome distribution of the condition with the higher selection rate accordingly. We then estimated lower and upper bounds by considering worst-case scenarios in which participants differentially selected into the analysis sample had either relatively high or relatively low outcome values. These analyses provide a conservative assessment of whether the direction of the estimated effects is robust to differential selection into the attention-check-passing sample.

The results in Table 39 indicate that several effects remain directionally robust even under these worst-case selection scenarios, including all effects identified in the main analysis. However, these results are not directly comparable to the estimates from our primary factorial models because the Lee-style bounds are implemented through binary treatment comparisons, each analysis compares a single treatment cell from the 2×2 design with the control condition, whereas our main analyses estimate the effects of risk, causal framing, and their interaction jointly across all four experimental conditions.

Even so, we found that among Democrats, the negative effects of the Risk-only and Causal-only conditions on Conserv+ topics remained negative across the full range of bounds, with the additional negative effect of the Risk-only condition on Budget and the negative effect of the Causal-only condition on AI Trust. Among Republicans, the negative effect of the Risk-only condition on budget remained negative across the bounds, with the additional negative effects of Risk on Liberal+ opinion and Risk + Causal on budget. In addition, the positive effect of the Risk-only condition on AI Trust remained positive throughout the bounds as well.

For the remaining comparisons, the bounds included zero, indicating that the direction of the estimated effect was not robust to the most extreme selection scenarios considered. Overall, these results suggest that some of the observed effects are robust to differential attention-check selection, while others are more sensitive to assumptions about the outcomes of participants whose inclusion in the analysis sample may have been affected by treatment assignment.

The Lee-style sensitivity analysis for the one-week follow-up outcomes similarly examined whether restricting the analyses to participants who passed the post-treatment attention check could affect the direction of the estimated treatment effects as seen in Table 40. Overall, several point estimates remained in the same direction across both extreme selection scenarios, although the majority of bounds included

zero. Among Democrats, the Risk + Causal condition remained negatively associated with both Conserv+ and Liberal+ outcomes across the full range of bounds, and the Causal-only condition remained negatively associated with AI Trust. Among Republicans, the Risk-only condition remained positively associated with Conserv+, Liberal+, Budget, and AI Trust, while the Causal-only condition remained positively associated with Budget. The remaining comparisons had bounds spanning zero, indicating that their estimated directions were more sensitive to the extreme assumptions about differential selection considered here. As these are Lee-style point-estimate bounds rather than confidence intervals, bounds that exclude zero should not be interpreted as evidence of statistical significance; instead, they indicate that the direction of the corresponding point estimate remains unchanged under the worst-case selection scenarios.

C.4.2 User Characteristics, AI Familiarity, and Post-Task AI Trust as Covariate. We also repeated the analysis from Section 5.3, Section 5.4, and Section 5.5 excluding those who failed the attention check. Again, we see minimal change to substantive results, except the user characteristics analysis.

Table 41: Personality Traits as Predictors of Persuasion and AI Trust (Pooled Model) - Excluded Participants

	Conserv+ Opinion	Liberal+ Opinion	Budget Allocation	AI Trust
Trait	*b* (SE)	*b* (SE)	*b* (SE)	*b* (SE)
Risk	−0.12(0.16)	−0.07(0.15)	−4.93** (1.82)	−0.03(0.13)
Causal	−0.20(0.16)	−0.08(0.14)	−1.59(1.88)	−0.15(0.13)
Risk×Causal	0.30(0.22)	0.16(0.21)	4.20(2.56)	0.14** (0.19)
Republican	**1.06**** (0.11)	**-0.54**** (0.11)	**6.09**** (1.50)	**0.72**** (0.10)
Extraversion	0.08* (0.04)	0.07(0.04)	0.03(0.44)	**0.09**** (0.03)
Openness	−0.09(0.05)	0.06(0.05)	0.08(0.68)	−0.01(0.04)
Conscientiousness	0.09(0.05)	−0.07(0.05)	0.18(0.64)	0.03(0.04)
Emotional Stability	0.02(0.04)	−0.10* (0.04)	−0.58(0.56)	0.04(0.04)
Agreeableness	−0.06(0.05)	0.09(0.05)	1.16* (0.56)	**0.14**** (0.04)
Self-Esteem	−0.02(0.06)	0.08(0.06)	0.18(0.73)	0.14* (0.06)

Note. Entries are OLS regression coefficients with HC3 robust standard errors in parentheses. Models are estimated on the pooled sample, controlling for party, disclaimer condition (Risk, Causal, Risk×Causal), and the corresponding pre-measure (opinion and budget models only). ** = $p_{\text{Holm}} < .05$, * = $p_{raw} < 0.05$.

Table 42: AI Familiarity Measures as Predictors of Persuasion and AI Trust (Pooled Model) - Excluded Participants

	Conserv+ Opinion	Liberal+ Opinion	Budget Allocation	AI Trust
Predictor	*b* (SE)	*b* (SE)	*b* (SE)	*b* (SE)
Risk	−0.13(0.16)	−0.05(0.15)	**-4.58**** (1.82)	−0.03(0.14)
Causal	−0.18(0.16)	0.00(0.15)	−0.83(1.84)	−0.04(0.14)
Risk × Causal	0.28(0.22)	0.08(0.21)	3.26(2.54)	0.04(0.19)
Republican	**1.11**** (0.11)	**-0.58**** (0.11)	**5.87**** (1.37)	**0.78**** (0.10)
AI Knowledge	**-0.21**** (0.09)	**-0.32**** (0.08)	**-3.49**** (1.09)	−0.14(0.08)
AI Education	0.14(0.10)	0.14(0.09)	0.88(1.32)	0.18* (0.09)
AI Use	**0.20**** (0.07)	**0.17**** (0.06)	**2.52**** (0.81)	**0.37**** (0.06)

Note. Entries are OLS regression coefficients with HC3 robust standard errors in parentheses. Models are estimated on the pooled sample, controlling for party, disclaimer condition (Risk, Causal, Risk×Causal), and the corresponding pre-measure (opinion and budget models only). This analysis is exploratory. ** = $p_{\text{Holm}} < .05$, * = $p_{raw} < 0.05$.

In Table 41, we present the results of the regression models including all user personality covariates. Interestingly, this analysis produces the largest change relative to our main results. All statistically significant effects of AI Trust remain unchanged; however, the effects of both Emotional Stability and Agreeableness are no longer statistically significant. This pattern may suggest a relationship between these personality characteristics and the participants who did not recall the attention warning presented at the end of the study. Although exploratory, this finding is interesting and may point to a promising avenue for future research.

Table 43: Effect of Disclaimer Conditions and AI Trust on Persuasion - Excluded Participants

		Conserv+ Opinion	Liberal+ Opinion	Budget Allocation
Party	Predictor	*b* (SE)	*b* (SE)	*b* (SE)
Democrat	Risk	−0.25(0.21)	0.02(0.21)	−2.08(2.08)
	Causal	−0.32*(0.21)	−0.04(0.22)	−0.19(2.24)
	Risk×Causal	0.40(0.28)	−0.12(0.30)	1.82(3.02)
	AI Trust	**0.45****(0.04)	**0.26****(0.05)	**4.77****(0.55)
Republican	Risk	0.03(0.22)	−0.16(0.20)	**-6.95****(2.84)
	Causal	0.03(0.22)	−0.01(0.19)	−1.66(2.71)
	Risk×Causal	0.11(0.30)	0.40(0.27)	5.61(3.91)
	AI Trust	**0.11****(0.05)	**0.30****(0.05)	**4.14****(0.63)

Note. Entries are OLS regression coefficients with HC3 robust standard errors in parentheses. Models are estimated separately by party. AI Trust is included as a covariate to same model used in H1 and H2. Hypothesis tests for disclaimer conditions (Risk, Causal, Risk×Causal) are one-tailed, directional negative; AI Trust is tested two-tailed. ** = $p < .05$; * = $p < .10$.

In Table 42 and Table 43, we present the remaining two sensitivity analyses. Results are largely consistent with the main analyses. The only difference is that the effect of AI Knowledge on post-task AI Trust is reduced from statistical significance at $p < .05$ to marginal significance at $p < .10$.

C.5 Pooled Models with Demographic Covariates

In this section, we provide pooled analyses for the models presented in Section 4.1, including all demographic variables that differed significantly between our two group variables: age, education, income, and race. Because these analyses were not hypothesized in our pre-registration, they should be interpreted as exploratory rather than causal analyses.

Table 44 present the results. Overall, the findings show no meaningful changes relative to the minimal models reported in the main paper.

Table 44: Effect of Disclaimer Conditions on Opinion, Budget Allocation, and AI Trust (Full Model) - All Participants

	Conserv+ Opinion	Liberal+ Opinion	Budget Allocation	AI Trust
Predictor	*b* (SE)	*b* (SE)	*b* (SE)	*b* (SE)
Risk	−0.15(0.13)	0.00(0.13)	**-3.20****(1.48)	0.03(0.11)
Causal	**-0.27****(0.13)	0.04(0.12)	−0.44(1.53)	0.00(0.11)
Risk×Causal	0.40(0.18)	−0.00(0.18)	2.78(2.17)	−0.04(0.16)
Republican	**1.07****(0.10)	**-0.53****(0.10)	**5.77****(1.17)	**0.88****(0.08)

Note. Entries are OLS regression coefficients with HC3 robust standard errors in parentheses. Models are estimated on the full sample; Republican (vs. Democrat) is included as a main-effect covariate only (no interactions with disclaimer conditions). Hypothesis tests for disclaimer main effects (Risk, Causal, Risk×Causal) are one-tailed, directional negative; Republican is tested two-tailed. Models for opinion and budget allocation include the corresponding pre-measure as a covariate (mean-centered); the trust model does not. ** = $p < .05$; * = $p < .10$.

C.6 AI Trust and Persuasion Regression Results

Interestingly, we found a strong trend between post-task AI Trust and the effect of persuasion of the model. To explore this, we added post-task AI Trust as a covariate to the H1 and H2 models, estimated separately by party. AI Trust was tested two-tailed as we had not pre-registered a direction; `Risk`, `Causal`, and their interaction remained one-tailed as in H1 and H2.

Results, seen in Table 45, show post-task AI Trust is a strong, consistent linear predictor of persuasion across all three outcomes and both parties. Among Democrats, higher post-task trust was associated with significantly larger pro-model shifts in all tasks (Conserv+ opinion: $b = 0.42$, Liberal+ opinion: $b = 0.25$, and budget allocation: $b = 5.06$; all $p < .05$). Among Republicans, the same held for all three outcomes (Conserv+ opinion: $b = 0.11$, Liberal+ opinion: $b = 0.31$, Budget: $b = 3.50$; all $p < .05$). Disclaimer effects were largely unchanged after controlling for AI trust, relative to H1 and H2, suggesting AI trust does not simply substitute for the disclaimer effects reported there.

Table 45: Effect of Disclaimer Conditions and AI Trust on Persuasion - All Participants

		Conserv+ Opinion	Liberal+ Opinion	Budget Allocation
Party	Predictor	b (SE)	b (SE)	b (SE)
Democrat	Risk	-0.24^{*} (0.18)	0.01 (0.18)	0.56 (1.81)
	Causal	**-0.29**** (0.17)	0.08 (0.18)	1.73 (1.99)
	Risk×Causal	0.48 (0.24)	−0.19 (0.26)	−0.61 (2.74)
	AI Trust	**0.42**** (0.04)	**0.25**** (0.04)	**5.06**** (0.47)
Republican	Risk	−0.04 (0.18)	−0.05 (0.17)	**-6.78**** (2.23)
	Causal	−0.18 (0.18)	−0.00 (0.16)	−2.11 (2.20)
	Risk×Causal	0.25 (0.26)	0.24 (0.23)	5.99 (3.21)
	AI Trust	**0.11**** (0.04)	**0.31**** (0.04)	**3.50**** (0.52)

Note. Entries are OLS regression coefficients with HC3 robust standard errors in parentheses. Models are estimated separately by party. AI Trust is included as a covariate in the same model used in H1 and H2. Hypothesis tests for disclaimer conditions (Risk, Causal, Risk×Causal) are one-tailed, directional negative; AI Trust is tested two-tailed. ** = $p < .05$; * = $p < .10$.

C.7 One Week Follow-Up Change Visualization

To better understand the persuasive effects over time, we include Figure 11 as a visualization of average opinion change per task between pre-task and post-task, compared to pre-task and 1-week follow up. Consistent with the patterns in our regression (Table 5, the mean change in Figure 11 generally remained in the direction of the immediate effects rather than returning fully to baseline a 0.

C.8 Sensitivity Checks

As mentioned in our pre-registration, we include a few sensitivity analyses below. All show strong robustness of results.

Follow-Up Analysis with Full Model. We present the H4 analysis using the full model, both as a minimal specification with no additional covariates and a specification that additionally controls for all demographic characteristics on which Democrats and Republicans who completed the follow-up differed significantly: age, education, Hispanic ethnicity, and race.

Interestingly, when pooling the data in the minimal model (Table 46), we observe a statistically significant decrease in persuasion from the pre-task measure to the one-week follow-up for the Causal condition in the Conserv+ task. However, this effect is reduced to marginal significance ($p < .10$) when demographic covariates are included (Table 47). Taken together, these results lend additional support to our hypothesis that the effects of the intervention diminish over time but do not fully reverse one week later.

Table 46: Follow-Up Sensitivity Check: Disclaimer Effects Controlling for Partisanship (Pre-Task, Minimal Model)- All Participants

	Conserv+ Opinion	Liberal+ Opinion	Budget Allocation
Predictor	b (SE)	b (SE)	b (SE)
Risk	−0.11 (0.14)	0.06 (0.15)	−0.71 (1.82)
Causal	**-0.24**** (0.14)	−0.04 (0.14)	1.35 (1.89)
Risk×Causal	0.11 (0.20)	−0.02 (0.20)	1.88 (2.72)
Republican	**0.52**** (0.10)	**0.55**** (0.11)	**3.57**** (1.42)

Note. Entries are OLS regression coefficients with HC3 robust standard errors in parentheses. Models are estimated on the full sample with Republican (vs. Democrat) included as a covariate. Pre-treatment values of the corresponding outcome are included as covariates. Hypothesis tests for disclaimer conditions (Risk, Causal, Risk×Causal) are one-tailed, directional negative; Republican is tested two-tailed. ** = $p < .05$; * = $p < .10$.

Analysis with Change Scores. Next, we re-analyze H1-H4 using the change in responses (y_{POST} - y_{PRE}) model interaction instead of regressing for y_{POST} using y_{PRE} as a covariate. Results are seen in Table 48 for the main study and Table 49 for the follow-up analysis. We find no difference in significance found compared to the results in our main study.

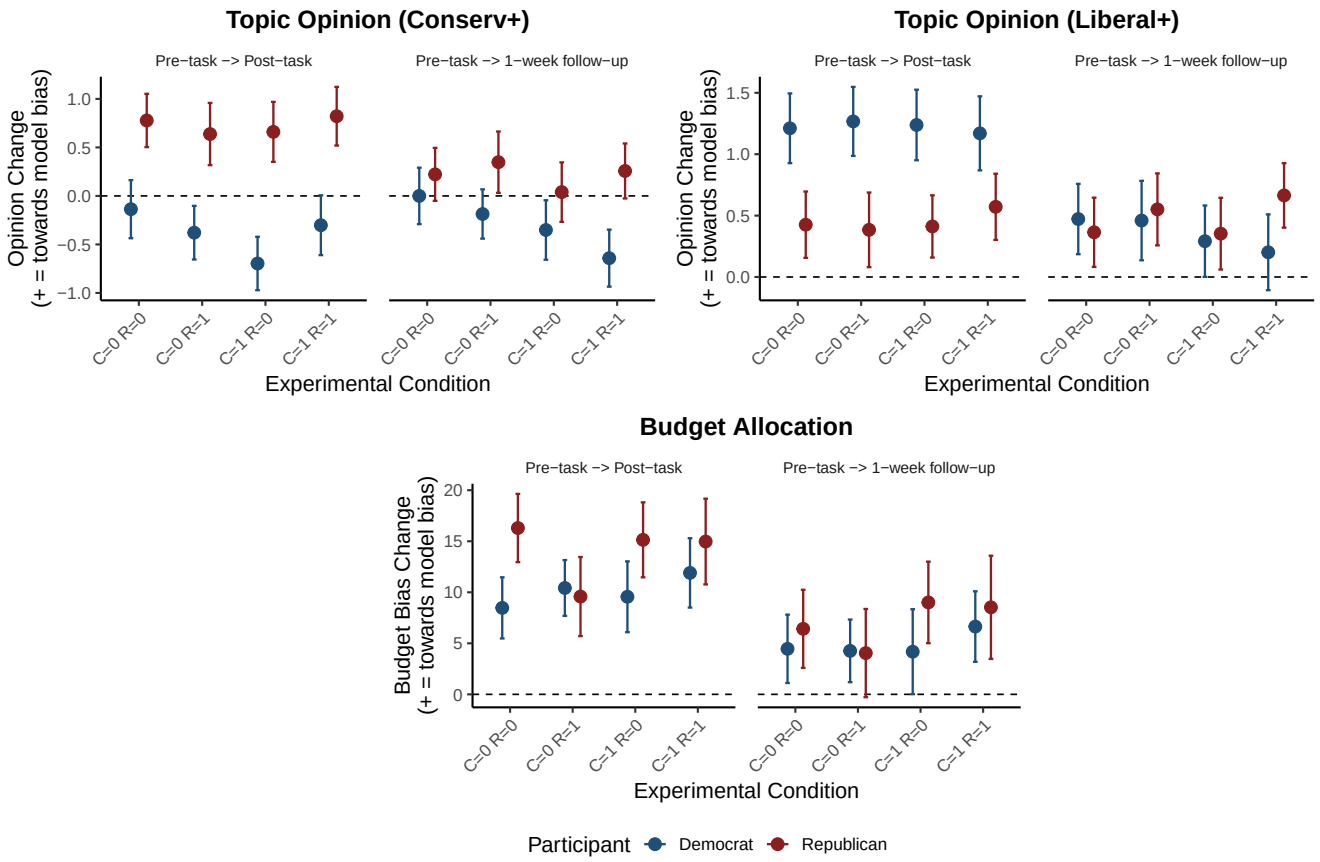


Figure 11: Mean change in outcomes from pre-task to post-task and pre-task to one-week follow-up. Points represent mean within-participant change, with 95% confidence intervals, separately for Democrats and Republicans across experimental conditions. Experimental conditions are indicated by `Causal`, "C" (0=absent, 1=present) and `Risk`, "R" (0=inaccuracy, 1=political bias). Positive values indicate movement toward the model's bias. The panels show changes in opinion for the Conservative+ and Liberal+ topic tasks and changes in budget allocations.

Analysis with Logistic Regression. Next, we re-estimated our primary models using ordinal logistic regression rather than ordinary least squares. The post-treatment outcomes and trust measures, which were originally analyzed as approximately continuous Likert-type scales, were instead treated as ordered categorical variables. Specifically, each outcome was converted to an ordered factor preserving its original response ordering, and proportional odds models were estimated using 'polr'. The same model specifications, including all covariates and interaction terms, were retained to assess whether our substantive conclusions were robust to alternative assumptions about the scale properties of the dependent variables. We again see no significant change in the follow-up results (Table 50, Table 51).

Given these results we feel confident in our analysis in the main paper.

Table 47: Follow-Up Sensitivity Check: Disclaimer Effects Controlling for Partisanship (Pre-Task, Full Model)- All Participants

	Conserv+ Opinion	Liberal+ Opinion	Budget Allocation
Predictor	*b* (SE)	*b* (SE)	*b* (SE)
Risk	−0.12(0.14)	0.07(0.15)	−0.63(1.82)
Causal	−0.25*(0.14)	−0.04(0.14)	1.31(1.88)
Risk×Causal	0.11(0.20)	−0.03(0.20)	1.79(2.72)
Republican	**0.52****(0.10)	**0.59****(0.12)	**4.15****(1.41)

Note. Entries are OLS regression coefficients with HC3 robust standard errors in parentheses. Models are estimated on the full sample with Republican (vs. Democrat) included as a covariate, along with age, education, and race covariates. Pre-task values of the corresponding outcome are included as covariates. Hypothesis tests for disclaimer conditions (Risk, Causal, Risk×Causal) are one-tailed, directional negative; Republican is tested two-tailed. ** = $p < .05$; * = $p < .10$.

Table 48: Main Study Sensitivity Check: Disclaimer Effects on Opinion and Budget Change Scores ($y_{\text{post}} - y_{\text{pre}}$) - All Participants

		Conserv+ Opinion	Liberal+ Opinion	Budget Allocation
Party	Predictor	*b* (SE)	*b* (SE)	*b* (SE)
	Risk	−0.18(0.19)	−0.03(0.19)	−0.38(1.91)
Democrat	Causal	**-0.39****(0.19)	0.03(0.19)	1.49(2.20)
	Risk×Causal	0.49(0.27)	−0.10(0.27)	0.56(3.01)
	Risk	0.03(0.19)	0.01(0.18)	**-6.74****(2.27)
Republican	Causal	−0.13(0.19)	−0.01(0.16)	−1.98(2.25)
	Risk×Causal	0.15(0.26)	0.16(0.24)	5.03(3.28)

Note. Entries are OLS regression coefficients with HC3 robust standard errors in parentheses. The dependent variable is the post-pre difference score rather than the ANCOVA specification used in primary analyses. All hypothesis tests are one-tailed, directional negative. AI Trust is excluded as it has no pre-measure. ** = $p < .05$; * = $p < .10$.

Table 49: Follow-Up Sensitivity Check: Disclaimer Effects on Opinion and Budget Change Scores ($y_{\text{fu}} - y_{\text{pre}}$)- All Participants

		Conserv+ Opinion	Liberal+ Opinion	Budget Allocation
Party	Predictor	*b* (SE)	*b* (SE)	*b* (SE)
	Risk	−0.19(0.20)	−0.01(0.22)	−0.20(2.33)
Democrat	Causal	−0.35*(0.22)	−0.18(0.21)	−0.28(2.73)
	Risk×Causal	−0.10(0.29)	−0.08(0.31)	2.66(3.62)
	Risk	0.13(0.21)	0.19(0.21)	−2.38(2.96)
Republican	Causal	−0.18(0.21)	−0.01(0.21)	2.59(2.83)
	Risk×Causal	0.09(0.30)	0.12(0.29)	1.90(4.43)

Note. Entries are OLS regression coefficients with HC3 robust standard errors in parentheses. The dependent variable is the follow-up minus pre-task difference score rather than the ANCOVA specification used in primary follow-up analyses. All hypothesis tests are one-tailed, directional negative. ** = $p < .05$; * = $p < .10$.

Table 50: Main Study Sensitivity Check: Ordinal Logistic Regression of Disclaimer Effects - All Participants

		Conserv+ Opinion	Liberal+ Opinion	Budget Allocation	AI Trust
Party	Predictor	*b* (SE)	*b* (SE)	*b* (SE)	*b* (SE)
Democrat	Risk	**-0.31**** (0.18)	0.02 (0.18)	−0.01 (0.17)	−0.10 (0.18)
	Causal	**-0.36**** (0.18)	0.06 (0.18)	0.08 (0.18)	−0.15 (0.17)
	Risk×Causal	0.60 (0.26)	−0.14 (0.25)	−0.02 (0.25)	0.34 (0.25)
Republican	Risk	−0.05 (0.18)	0.03 (0.18)	**-0.38**** (0.18)	0.17 (0.17)
	Causal	−0.22 (0.18)	0.05 (0.18)	−0.14 (0.17)	0.16 (0.17)
	Risk×Causal	0.31 (0.25)	0.05 (0.26)	0.23 (0.25)	**-0.44**** (0.25)

Note. Entries are log-odds coefficients from proportional odds (ordinal logistic) regression models estimated via `polr`, with standard errors in parentheses. Post-treatment outcomes and trust measures were treated as ordered factors. The same model specifications as the primary analyses were retained. All hypothesis tests are one-tailed, directional negative. ** = $p < .05$; * = $p < .10$.

Table 51: Follow-Up Sensitivity Check: Ordinal Logistic Regression of Disclaimer Effects- All Participants

		Conserv+ Opinion	Liberal+ Opinion	Budget Allocation
Party	Predictor	*b* (SE)	*b* (SE)	*b* (SE)
Democrat	Risk	−0.29* (0.20)	0.07 (0.20)	−0.07 (0.20)
	Causal	−0.35** (0.20)	−0.14 (0.19)	−0.18 (0.20)
	Risk×Causal	−0.01 (0.28)	−0.16 (0.28)	0.26 (0.28)
Republican	Risk	0.08 (0.20)	0.09 (0.21)	−0.08 (0.20)
	Causal	−0.23 (0.20)	0.10 (0.20)	0.22 (0.20)
	Risk×Causal	0.15 (0.29)	0.10 (0.29)	0.11 (0.29)

Note. Entries are log-odds coefficients from proportional odds (ordinal logistic) regression models estimated via `polr`, with standard errors in parentheses. Follow-up outcomes were treated as ordered factors; post-treatment values of the corresponding outcome were included as covariates. All hypothesis tests are one-tailed, directional negative. ** = $p < .05$; * = $p < .10$.