AI-Generated Content and Public Persuasion: The Limited Effect of AI Authorship Labels

Isabel O. Gallegos[1,2], Chen Shani[1], Weiyan Shi[*,3], Federico Bianchi[1], Robb Willer[4], Dan Jurafsky[1]
[1]Department of Computer Science, Stanford University [2]Stanford Law School, Stanford University
[3]Department of Computer Science, Northeastern University [4]Department of Sociology, Stanford University

Abstract

As the growing capabilities of generative AI models have enabled information creation and dissemination at increased scale and speed, labeling AI-generated content has emerged as a policy proposal to increase transparency to users and reduce risks of misinformation and deception. While prior work has investigated the persuasiveness of AI-generated political messaging without notifying people of the information author, it is still unclear whether the authorship of information affects its persuasiveness. It is critical to understand this impact, given the proposed policy regarding labeling AI-generated content. In this study, we conduct a survey of U.S. respondents to investigate the persuasiveness of information labeled as AI-generated, human-written, or unlabeled across four policy issues. We find that the disclosure of AI authorship does not significantly affect persuasiveness. Further, this relationship holds even when controlling for respondents' prior knowledge about the policy, political party, education level, age, and prior experience with AI tools. These results suggest that labeling content as AI-generated may have only a weak effect and may minimally diminish its persuasive impact. Thus, this calls for other solutions to address the pressing issue.

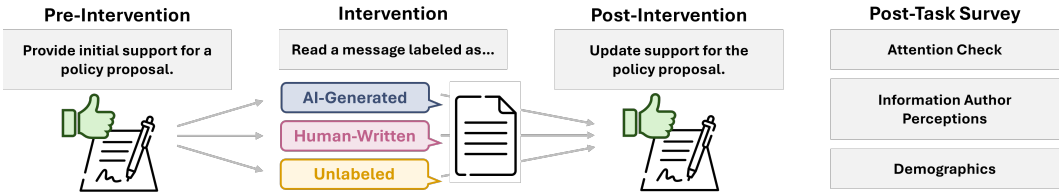

Figure 1: In this study, we measure the persuasiveness of information with different authorship labels across four different policy proposals. We find that the authorship label of the message as AI-generated, human-written, or unlabeled does not significantly affect its persuasiveness.

1 Introduction

Generative artificial intelligence (AI) tools can now write persuasive content with scale and speed that exceeds human-written information [2, 9, 12, 16]. With these advancing capabilities have emerged new concerns about AI-powered influence operations [16, 15], misinformation campaigns [27], and other deceptive political activities. Even when not intentionally misleading, AI-generated political content can exacerbate bias [5] and disrupt the information ecosystem with personalized, targeted materials that require no professional expertise or manual effort [12]. Complicating matters, people

[*]Work completed at Stanford University.

2nd Workshop on Regulatable ML at NeurIPS 2024.

struggle to detect AI-generated content, often using flawed heuristics [23], leading to an inability to distinguish it from human-written messages [7, 34].

A common policy response has been to identify AI-generated content with an authorship label [41]. U.S. Executive Order 14110 on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence [11], for instance, calls for "steps to watermark or otherwise label output from generative AI," with companies like OpenAI, Alphabet, and Meta committing to such measures [3]. The EU AI Act requires that deployers of AI-generated or manipulated content provide AI disclosure labels, and the AI Labeling Act of 2023 [38], introduced in the U.S. Senate, and the AI Disclosure Act of 2023 [37], introduced in the U.S. House of Representatives, call for similar provisions. Techniques like generative text watermarking may enable persistent traceability for such disclosure requirements [8].

Still, the influx of these labeling policy proposals opens up a critical question: **do authorship labels change people's utilization of information?** In this study, we examine political persuasion, one of the most pressing concerns in policy debates around AI's ability to influence people's political beliefs, attitudes, and behaviors. In a pre-registered survey experiment, we test the extent to which different authorship labels change participants' level of support for four different policy proposals. To do so, we compare the persuasive effects of messages labeled as AI-generated to messages labeled as human-written or unlabeled, keeping the message content fixed and manipulating only the authorship disclosure. We seek to answer the following research questions:

- **RQ1:** Are messages labeled as human-written more persuasive than those labeled as AI-generated?
- **RQ2:** Are messages labeled as AI-generated more persuasive than unlabeled messages?
- **RQ3:** How do AI authorship labels affect people's level of confidence in their support, sharing intentions of information, and accuracy judgments of information?

We hypothesize that messages labeled as AI-generated will be less persuasive than those labeled as human-written, and that messages labeled as AI-generated will be more persuasive than messages with no label.

While prior work has focused on either the persuasiveness of AI-generated content *without* authorship labels [2, 9, 14, 18, 19, 31], or the *perceptions* of labeled information such as credibility, reliability, or quality [1, 4, 6, 17, 28, 20–22, 24–26, 29, 30, 33, 36, 39, 43], our contribution lies in investigating the *utilization* – not just perception – of labeled information, namely the persuasiveness of information under source identification, given the same content. **We find that the notification of AI authorship does not significantly affect the persuasiveness of policy proposal messages, but may potentially have a small effect**. These results suggest a possible inefficacy of authorship labels to meaningfully change perceptions of information content, and highlight a need for both further investigation of labeling policies, as well as more thorough policy interventions that can have stronger mitigations against AI-powered information campaigns.

2 Methodology

2.1 Experimental Design

In our study, we measure the persuasiveness of information with different authorship labels across four policy proposals. Following the study design of Bai et al. [2], we measure persuasiveness by the change in support before and after reading a persuasive message. The study has a 2 (Time: Pre-intervention vs. Post-intervention) x 4 (Policy: Geoengineering vs. Drug Importation vs. College Athlete Salaries vs. Social Media Platform Liability) x 3 (Authorship Label: AI Label vs. Human Label vs. No Label) within-between-between-subject design. As illustrated in Figure 1, participants read the exact same message about their assigned policy but are either told that it is generated by an expert AI model, written by a policy expert, or given no authorship details. The No Label condition represents a scenario where the information source is unspecified, a common instance in online content, while the Human Label condition provides a direct contrast between two known sources.

We randomly sampled four policy proposals from Durmus et al. [9]'s Persuasion Dataset (CC BY-NC-SA 4.0), which contains a set of 56 claims about emerging issues, such as geoengineering and collegiate athlete salaries. Given the developing nature of the issues, people's opinions are expected to be less entrenched and more impressionable to persuasive arguments. In June 2024, we collected one message generated by OpenAI's GPT-4o model for each of the four policy proposals to persuade

readers to support the policies, controlling for the persuasiveness technique [42]. We manually edited the text only to correct any factual errors. For details, see Appendix A.

The study procedure has four stages. (1) In the pre-intervention stage, participants are told that they are participating in a public opinion survey to ground their responses in a realistic scenario; they then indicate their prior knowledge about the topic and provide the pre-intervention measures of support and confidence. (2) In the intervention stage, participants receive information, labeled according to their assigned authorship label condition. (3) In the post-intervention stage, participants provide the post-intervention measure of support, confidence, sharing intention, and accuracy judgment. (4) Finally, participants complete a post-task survey that asks for an attention check, their perceptions of their assigned information source, and their demographic information. At the end of the survey, we disclose the purpose of the study and the true authorship of the messages. The study was approved by Stanford University's Institutional Review Board, and all participants provided and confirmed their informed consent.

With support as the primary measure of persuasion, we measure the following dependent variables on a 0-100 scale, following Bai et al. [2]. For the complete set of measures, see Appendix B.

- **Support.** Please indicate your level of agreement with the following statement: [*Policy Proposal*].
- **Confidence.** How confident are you in your responses?
- **Sharing Intention.** How likely would you be to share this information with others if the topic came up (for example, in conversation or on social media)?
- **Accuracy Judgement.** To the best of your knowledge, is the information accurate?

We conducted an *a priori* power analysis based on pilot data to determine the required sample size. Assuming an effect size of 3 points on the raw 0-100 point scale for the support variable between the AI Label and Human Label conditions, a sample size of 1,500 was required to achieve a power of 80 at a significance level of 0.05; though prior work has not examined persuasiveness of AI versus human authorship labels, this effect size is based on prior literature that has found very small effects between AI and human authorship disclosures along other perception measures [20, 24, 30, 33, 36]. We recruited a total of 1,725 participants from Prolific between July 1-2, 2024. We excluded participants who failed the attention check question, as well as participants who failed two manipulation check questions that checked if they believed the author was the same as their assigned condition. This exclusion ensured that we retained only those participants who were successfully influenced by the manipulation. Our final sample was 1,515 participants. For participant details, see Appendix C.

2.2 Analysis

We estimate several regression models, following the method of Bai et al. [2], to model multiple control variables and interaction effects. To answer RQ1, we regress the post-intervention dependent variable of support on the dummy-coded variables for the AI Label and No Label conditions (contrasted with the Human Label condition) while controlling for the policy (dummy-coded, contrasted with Geoengineering); we additionally control for the pre-intervention measure of support. We consider all policies within a single regression model (as opposed to modeling each policy condition separately) following Fong and Grimmer [13]'s recommendation and Bai et al. [2]'s methodology to improve causal inference compared to a single policy. To address RQ2, we again regress the post-intervention dependent variable of support on the dummy-coded variables for the AI Label and Human Label conditions (contrasted with the No Label condition), again controlling for the policy and the pre-intervention measure.

We additionally perform two equivalence tests with two one-sided tests (TOST) with Welch's test statistic, which assumes unequal variance [40]. In the first TOST equivalence test, we compare the mean difference between the post-intervention and the pre-intervention support variables in the AI Label condition to that of the Human Label condition to test the alternative hypothesis that the means do not differ by more than a small amount, defined by the equivalence bounds. In the second TOST equivalence test, we instead compare the means between the AI Label and the No Label condition. These tests allow us to understand within what interval the differences between conditions lie.

Finally, to understand RQ3, we repeat the regression analyses used for RQ1 and RQ2, but interchange the dependent variable of support. We only control for the pre-intervention measure when the dependent variable is confidence.

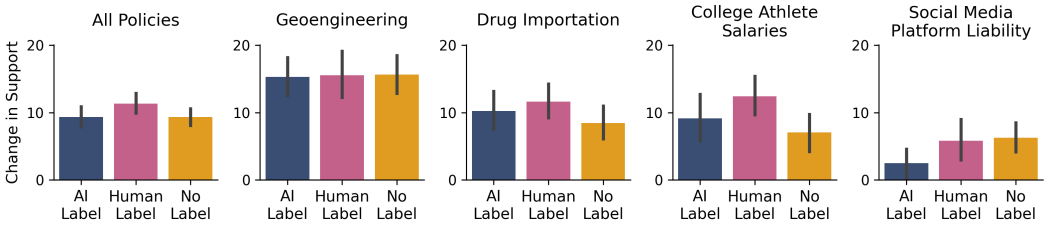

Figure 2: Mean and 95% confidence intervals for the difference between the post-intervention and pre-intervention support variables for the AI Label, Human Label, and No Label conditions.

3 Results

Across all policies, we find no evidence that AI authorship notifications significantly change the persuasiveness of the political messaging. We show the mean differences and 95% confidence intervals between the post-intervention and pre-intervention support variables across all policies in Figure 2. Turning first to RQ1, we find that, even though participants assigned to the Human Label condition tended to increase their support more than participants assigned to the AI Label condition across every policy, the difference between conditions was not significant when analyzing our combined regression model that includes all topics (b=-1.82, CI=[-3.74, 0.10], p=0.063), nor when analyzing each topic individually. Comparing the AI Label condition and the No Label condition for RQ2, we again see no statistically significant difference (b=-0.088, CI=[-1.96, 1.78], p=0.92). Finally, for RQ3, we do not observe any significant differences between labeling conditions for the confidence, sharing, and accuracy dependent variables. For full regression results, see Appendix D.

Given the similarity in the persuasiveness of the messages across the conditions, we use the TOST equivalence tests to determine the equivalence bounds at which the authorship labeling conditions are statistically equivalent with respect to the difference between the post-intervention and pre-intervention support variables, illustrated in Figure 10 in Appendix D. We find that the AI Label condition is statistically equivalent to the Human Label condition at equivalence bounds of 3.86 points or more on the 0-100 scale, and is equivalent to the No Label condition at 1.68 points or more.

We also examine if other variables moderate the persuasive effect of each of the authorship label conditions by separately adding interaction terms between the pre-intervention support and each of the dummy variables for the following possible moderators: political party identity, prior knowledge about the topic, prior experience with AI tools, education level, and age. We find no evidence that any of these variables moderated the effect of the authorship label.

These results suggest that authorship labels may not meaningfully change the persuasiveness of political information content, even across several potential moderating factors. These findings come in light of strong policy pushes to require such labels to change information perceptions and point to a need for stronger mitigations against AI-generated information, beyond an authorship label alone.

4 Limitations

Even though we do not find significant effects between labeling conditions, we do observe that the Human Label condition tends to be more persuasive than the AI Label condition. Despite conducting an *a priori* power analysis based on pilot data, it is possible that our sample size was unable to detect this small effect. Future work can build on our results to better characterize this difference. However, even if significant, such a small effect likely leaves the policy implications unchanged. This work also only investigates one setting of persuasion, which simplifies theories that suggest that people leverage many heuristics beyond the information content alone to make judgments [32, 35]. Additionally, we consider a small number of messages with only three possible labels for information, but the possibilities of authorship notification can be much broader [10]. Future work can consider other content types and misinformation settings. Finally, we do not consider people's acceptance or aversion to AI *in general*, which may be an important moderating factor. Ultimately, trust in AI information sources is context-dependent [6], and persuasiveness of AI-mediated information may be no different.

5 Conclusions

We have presented one of the first investigations of the impact of AI disclosures on the persuasiveness of information and have shown that **the persuasive effect is nearly equivalent when information is labeled as AI-generated as when it is labeled as human-written or has no label**. Given the growing call for AI disclosures in emerging AI regulation and governance, these findings suggest that these policies may only have a weak effect on people's perceptions and utilization of labeled content. **These findings highlight a need for further investigation into the efficacy of AI disclosure policies**, while also emphasizing the importance of alternative and additional interventions, including media literacy education for the public, more informative cues about the trustworthiness and reliability of AI-generated content, and deamplification of AI-generated content when appropriate [12].

Acknowledgments

We thank Josh A. Goldstein and Jan G. Voelkel for helpful feedback. Isabel O. Gallegos is supported by the Fannie & John Hertz Foundation and Stanford Knight-Hennessy Scholars graduate fellowship.

References

[1] Sacha Altay and Fabrizio Gilardi. 2023. Headlines Labeled as AI-Generated Are Distrusted, Even When True or Human-Made, Because People Assume Full AI Automation. https://doi.org/10.31234/osf.io/83k9r

[2] Hui Bai, Jan G Voelkel, johannes C Eichstaedt, and Robb Willer. 2023. Artificial Intelligence Can Persuade Humans on Political Issues. https://doi.org/10.31219/osf.io/stakv

[3] Diane Bartz and Krystal Hu. 2023. OpenAI, Google, others pledge to watermark AI content for safety, White House says. https://www.reuters.com/technology/openai-google-others-pledge-watermark-ai-content-safety-white-house-2023-07-21/

[4] Amirsiavosh Bashardoust, Stefan Feuerriegel, and Yash Raj Shrestha. 2024. Comparing the willingness to share for human-generated vs. AI-generated fake news. arXiv:2402.07395 [cs.SI] https://arxiv.org/abs/2402.07395

[5] Scott Babwah Brennen and Matt Perault. 2024. *The New Political Ad Machine: Policy Frameworks for Political Ads in an Age of AI*. Technical Report. Center on Technology Policy at the University of North Carolina at Chapel Hill. https://techpolicy.unc.edu/wp-content/uploads/2023/11/GAI-and-political-ads.pdf

[6] Joy Buchanan and William Hickman. 2024. Do people trust humans more than ChatGPT? *Journal of Behavioral and Experimental Economics* 112 (2024), 102239. https://doi.org/10.1016/j.socec.2024.102239

[7] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 7282–7296. https://doi.org/10.18653/v1/2021.acl-long.565

[8] Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, et al. 2024. Scalable watermarking for identifying large language model outputs. *Nature* 634, 8035 (2024), 818–823.

[9] Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. 2024. *Measuring the Persuasiveness of Language Models*. Anthropic. https://www.anthropic.com/news/measuring-model-persuasiveness

[10] Ziv Epstein, Mengying C Fang, Antonio A Arechar, and David G Rand. 2023. What label should be applied to content produced by generative AI? https://doi.org/10.31234/osf.io/v4mfz

[11] Exec. Order. 14110. 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

[12] Stefan Feuerriegel, Renée DiResta, Josh A Goldstein, Srijan Kumar, Philipp Lorenz-Spreen, Michael Tomz, and Nicolas Pröllochs. 2023. Research can help to tackle AI-generated disinformation. *Nature Human Behaviour* 7, 11 (2023), 1818–1821.

[13] Christian Fong and Justin Grimmer. 2023. Causal inference with latent treatments. *American Journal of Political Science* 67, 2 (2023), 374–389.

[14] Josh A Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. 2024. How persuasive is AI-generated propaganda? *PNAS Nexus* 3, 2 (Feb 2024), pgae034. https://doi.org/10.1093/pnasnexus/pgae034 arXiv:https://academic.oup.com/pnasnexus/article-pdf/3/2/pgae034/56712546/pgae034.pdf

[15] Josh A Goldstein and Girish Sastry. 2023. The coming age of AI-powered propaganda. *Foreign Affairs* 27 (2023). https://www.foreignaffairs.com/united-states/coming-age-ai-powered-propaganda

[16] Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. arXiv:2301.04246 [cs.CY] https://arxiv.org/abs/2301.04246

[17] Andreas Graefe and Nina Bohlken. 2020. Automated journalism: A meta-analysis of readers' perceptions of human-written in comparison to automated news. *Media and Communication* 8, 3 (2020), 50–59.

[18] Kobi Hackenburg, Lujain Ibrahim, Ben M Tappin, and Manos Tsakiris. 2023. Comparing the persuasiveness of role-playing large language models and human experts on polarized U.S. political issues. https://doi.org/10.31219/osf.io/ey8db

[19] Kobi Hackenburg and Helen Margetts. 2024. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences* 121, 24 (2024), e2403116121. https://doi.org/10.1073/pnas.2403116121 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2403116121

[20] Lennart Hofeditz, Milad Mirbabaie, Jasmin Holstein, and Stefan Stieglitz. 2021. Do You Trust an AI-journalist? A Credibility Analysis of News Content with AI-Authorship. In *Twenty-Ninth European Conference on Information Systems (ECIS 2021)*. European Conference on Information Systems, Marrakech, Morocco, 1–15.

[21] Martin Huschens, Martin Briesch, Dominik Sobania, and Franz Rothlauf. 2023. Do You Trust ChatGPT? – Perceived Credibility of Human and AI-Generated Content. arXiv:2309.02524 [cs.HC] https://arxiv.org/abs/2309.02524

[22] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T. Hancock, and Mor Naaman. 2019. AI-Mediated Communication: How the Perception that Profile Text was Written by AI Affects Trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300469

[23] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences* 120, 11 (2023), e2208839120. https://doi.org/10.1073/pnas.2208839120 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2208839120

[24] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T. Hancock. 2023. Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 116 (apr 2023), 29 pages. https://doi.org/10.1145/3579592

[25] Castulus Kolo, Joschka Mutterlein, and Sarah Anna Schmid. 2022. Believing Journalists, AI, or Fake News: The Role of Trust in Media. In *Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS)*. Hawaii International Conference on System Sciences, Maui, Hawaii, USA, 1–10.

[26] Sarah Kreps, R. Miles McCain, and Miles Brundage. 2022. All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. *Journal of Experimental Political Science* 9, 1 (2022), 104–117. https://doi.org/10.1017/XPS.2020.37

[27] Christina LaChapelle and Catherine Tucker. 2023. Generative AI in Political Advertising. https://www.brennancenter.org/our-work/research-reports/generative-ai-political-advertising

[28] Angelica Lermann Henestrosa, Hannah Greving, and Joachim Kimmerle. 2023. Automated journalism: The effects of AI authorship and evaluative information on the perception of a science journalism article. *Computers in Human Behavior* 138 (2023), 107445. https://doi.org/10.1016/j.chb.2022.107445

[29] Yihe Liu, Anushk Mittal, Diyi Yang, and Amy Bruckman. 2022. Will AI Console Me when I Lose my Pet? Understanding Perceptions of AI-Mediated Email Writing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 474, 13 pages. https://doi.org/10.1145/3491102.3517731

[30] Chiara Longoni, Andrey Fradkin, Luca Cian, and Gordon Pennycook. 2022. News from Generative Artificial Intelligence Is Believed Less. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 97–106. https://doi.org/10.1145/3531146.3533077

[31] S. C. Matz, J. D. Teeny, S. S. Vaid, H. Peters, G. M. Harari, and M. Cerf. 2024. The potential of generative AI for personalized persuasion at scale. *Scientific Reports* 14, 1 (Feb 2024), 4692. https://doi.org/10.1038/s41598-024-53755-0

[32] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '94)*. Association for Computing Machinery, New York, NY, USA, 72–78. https://doi.org/10.1145/191666.191703

[33] Irene Rae. 2024. The Effects of Perceived AI Use On Content Perceptions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 978, 14 pages. https://doi.org/10.1145/3613904.3642076

[34] Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. AI model GPT-3 (dis)informs us better than humans. *Science Advances* 9, 26 (2023), eadh1850. https://doi.org/10.1126/sciadv.adh1850 arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.adh1850

[35] S. Shyam Sundar. 2008. The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility. In *Digital Media, Youth, and Credibility*, Miriam J. Metzger and Andrew J. Flanagin (Eds.). The MIT Press, Cambridge, MA, 73–100. https://doi.org/10.1162/dmal.9780262562324.073

[36] Benjamin Toff and Felix M Simon. 2023. "Or they could just not use it?": The Paradox of AI Disclosure for Audience Trust in News. https://doi.org/10.31235/osf.io/mdvak

[37] U.S. Congress. 2023. H.R.3831 - 118th Congress (2023-2024): AI Disclosure Act of 2023. https://www.congress.gov/bill/118th-congress/house-bill/3831/text

[38] U.S. Congress. 2023. S.2691 - 118th Congress (2023-2024): AI Labeling Act of 2023. https://www.congress.gov/bill/118th-congress/senate-bill/2691/text

[39] T. Franklin Waddell. 2018. A Robot Wrote This? How perceived machine authorship affects news credibility. *Digital Journalism* 6, 2 (2018), 236–255. https://doi.org/10.1080/21670811.2017.1384319 arXiv:https://doi.org/10.1080/21670811.2017.1384319

[40] Bernard L Welch. 1938. The significance of the difference between two means when the population variances are unequal. *Biometrika* 29, 3/4 (1938), 350–362.

[41] Chloe Wittenberg, Ziv Epstein, Adam J. Berinsky, and David G. Rand. 2024. Labeling AI-Generated Content: Promises, Perils, and Future Directions. https://mit-genai.pubpub.org/pub/hu71se89 Published in An MIT Exploration of Generative AI.

[42] Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How Johnny Can Persuade LLMs to Jailbreak Them: Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs. arXiv:2401.06373 [cs.CL] https://arxiv.org/abs/2401.06373

[43] Yunhao Zhang and Renée Gosline. 2023. Human favoritism, not AI aversion: People's perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation. *Judgment and Decision Making* 18 (2023), e41. https://doi.org/10.1017/jdm.2023.37

A Information Content

We use the following prompt to generate the information content from GPT-4o, employing techniques of evidence-based persuasion and expert endorsement described in Zeng et al. [42]'s taxonomy to control for persuasiveness:

```
Persuade a reader to agree with the following proposal: "[POLICY]". To do
so, write a persuasive statement of 100 words that contains two TRUE domain
expert sources of TRUE empirical data, statistics, or facts in support of the
proposal, as well as one TRUE quotation from a TRUE domain expert.
```

The generated information contents, with factual errors manually corrected, are shown in Table 1.

Geoengineering poses too many risks and should not be considered. Geoengineering poses too many risks and should not be considered. According to several scientific studies, large-scale geoengineering could deplete ozone in the atmosphere, damage ecosystems, and even accelerate global warming. Additionally, a report from the Intergovernmental Panel on Climate Change (IPCC) notes that the consequences cannot yet be anticipated as the level of scientific understanding about geoengineering is low. Dr. Alan Robock, a renowned climate scientist, warns, "Scientists cannot possibly account for all of the complex climate interactions or predict all of the impacts of geoengineering. With so much at stake, there is reason to worry about what we don't know." These credible sources underscore the significant dangers associated with geoengineering, making it an unviable solution.
Drug importation jeopardizes safety controls and the domestic pharma industry. Drug importation jeopardizes safety controls and the domestic pharma industry. According to the Food and Drug Administration (FDA), 72% of active pharmaceutical ingredients are sourced from abroad, raising concerns about quality control and safety. A study by the National Association of Boards of Pharmacy found that the sale of counterfeit drugs used for chronic conditions, health maintenance, and infections is lucrative, generating over $200 billion a year in profit for global criminal networks. Dr. Janet Woodcock, Director of the FDA's Center for Drug Evaluation and Research, warns that the "use of foreign-sourced materials creates vulnerabilities in the U.S. drug supply." Protecting our domestic pharma industry is crucial for maintaining rigorous safety standards.
College athletes should be paid salaries. College athletes should be paid salaries because they generate significant revenue for their institutions. According to the NCAA, college sports brought in $18.9 billion in 2019 alone. Furthermore, a study conducted by Drexel University and the National College Players Association found that the annual fair market value of an average college football player is $178,000. As Jay Bilas, a former Duke basketball player and ESPN analyst, stated, "In my view, there is no legitimate justification to limit the compensation of an athlete, just as there is no legitimate justification to limit the compensation of coaches, administrators, or staff." Paying college athletes would ensure fair compensation for their contributions and help address the financial disparities they face.
Social media platforms should be liable for harmful content posted by users. Social media platforms should be liable for harmful content posted by users. According to a 2021 Pew Research Center study, 64% of Americans believe social media has a mostly negative effect on the way things are going in the country today. Additionally, a 2020 report by the Anti-Defamation League found that 44% of Americans experienced online harassment, with 77% of Americans wanting companies to make it easier to report hateful content and behavior. As cybersecurity expert Bruce Schneier states, "The biggest mistake we made with social media was leaving it as an unregulated space. Even now – after all the studies and revelations of social media's negative effects – social media in the US remains largely an unregulated 'weapon of mass destruction."'

Table 1: Policy proposals and information content. Each of these employs the persuasion strategies of evidence-based persuasion and expert endorsement, taxonomized in Zeng et al. [42], in order to control for the level and type of persuasiveness.

B Measures

In the following survey questions, [*Policy Proposal*] denotes one of the policy proposals listed in Table 1, while [*Topic*] denotes one of the following phrases, based on the participants' assigned conditions: geoengineering, drug importation, college athlete salaries, or social media platform liability.

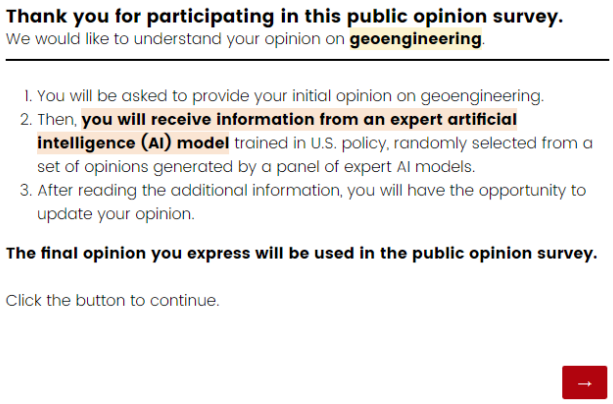

Figure 3: Example pre-intervention instructions for a participant assigned to the AI Label and Geoengineering conditions.

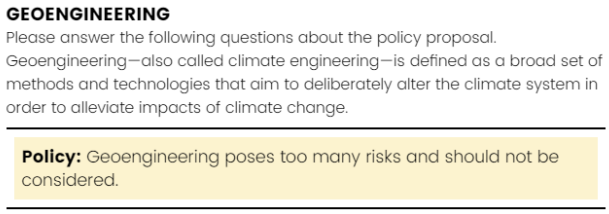

Figure 4: Example pre-intervention policy information for a participant assigned to the AI Label and Geoengineering conditions.

B.1 Pre-Intervention

Participants are first presented with instructions shown in Figure 3, and then are presented with a policy, as shown in Figure 4.

1. How knowledgeable do you feel about [*Topic*]?
   - I am an expert on this topic.
   - I am very knowledgeable about this topic.
   - I am moderately knowledgeable about this topic.
   - I am slightly knowledgeable about this topic.
   - I have little to no knowledge about this topic.
2. Please indicate your level of agreement with the following statement: [*Policy Proposal*].
   - 0 = Strongly disagree
   - 50 = Neither agree nor disagree
   - 100 = Strongly agree
3. How confident are you in your responses?
   - 0 = Very unconfident
   - 50 = Neither confident nor unconfident
   - 100 = Very confident

B.2 Intervention

In the AI Label condition, participants receive the following, with examples shown in Figures 5 and 6:

1. You will now receive information generated by an expert AI model trained in U.S. policy. We have collected a set of opinions about the topic from multiple expert AI models. Select the AI model that will provide the perspective you will read. (Participants select a number from 1 to 10.)
2. Consider the following information provided by expert AI model #[*Selection*]. [*Information Content*]

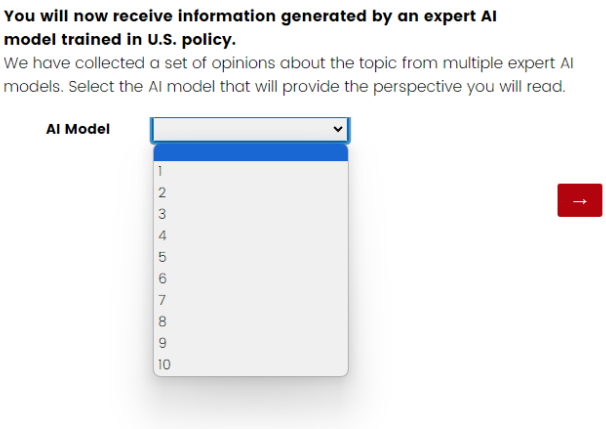

Figure 5: Example intervention instructions for a participant assigned to the AI Label and Geoengineering conditions.

☐ I have read this information carefully.

In the Human Label condition, participants receive the following:

1. You will now receive information generated by a policy expert trained in U.S. policy. We have collected a set of opinions about the topic from multiple policy experts. Select the person who will provide the perspective you will read. (Participants select a number from 1 to 10.)
2. Consider the following information provided by policy expert #[*Selection*]. [*Information Content*]

   ☐ I have read this information carefully.

In the No Label condition, participants receive the following:

1. You will now receive information. We have collected a set of opinions about the topic. Select the message that you will read. (Participants select a number from 1 to 10.)
2. Consider the following information provided by Message #[*Selection*]. [*Information Content*]

   ☐ I have read this information carefully.

B.3 Post-Intervention

After reading the information, participants respond to the following:

1. Please indicate your level of agreement with the following statement: [*Policy Proposal*]. You previously selected [*Pre-Intervention Selection*].
   - 0 = Strongly disagree
   - 50 = Neither agree nor disagree
   - 100 = Strongly agree
2. How confident are you in your responses? You previously selected [*Pre-Intervention Selection*].
   - 0 = Very unconfident
   - 50 = Neither confident nor unconfident
   - 100 = Very confident
3. How likely would you be to share this information with others if the topic came up (for example, in conversation or on social media)?
   - 0 = Very unlikely
   - 50 = Neither likely nor unlikely
   - 100 = Very likely
4. To the best of your knowledge, is the information accurate?
   - 0 = Very inaccurate

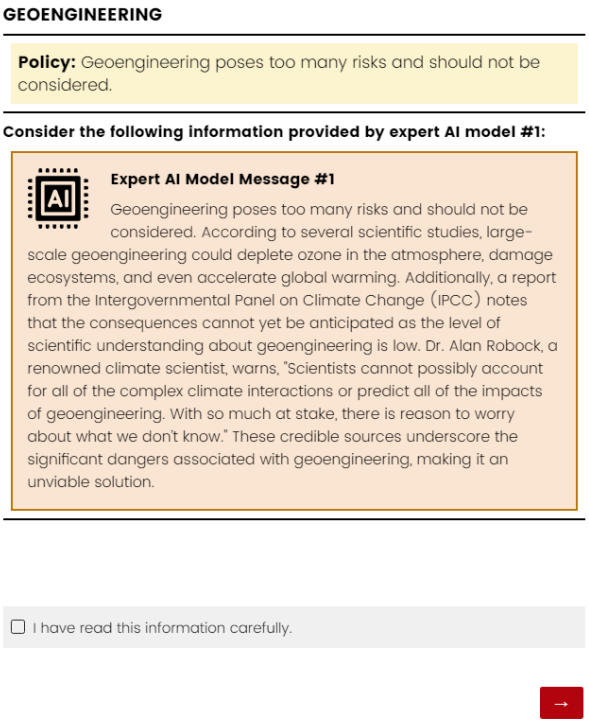

Figure 6: Example intervention message for a participant assigned to the AI Label and Geoengineering conditions.

- 50 = Neither accurate nor inaccurate
- 100 = Very accurate

B.4 Post-Task

Participants pass the attention check if they select the first and third answers:

> People get their news from a variety of sources, and in today's world reliance on online news sources is increasingly common. We also want to know if people are paying attention to the question. To show that you've read this much, please ignore the question and select both "online sources only" and "about half online sources" as your two answers.
>
> About how much of your news consumption comes from online sources? Please include print newspapers that you read online (e.g., washingtonpost.com) as online sources.
>
> - Online sources only
> - Mostly online sources
> - About half online sources
> - Mostly television or print news, with some online sources
> - Television or print news only

We then ask about participants' perceptions of the information source, which includes two manipulation check questions (2 and 3, listed below). Participants pass the manipulation check if they are assigned to the AI Label and Human Label condition and believe the author is the same as their assigned condition or do not question the author's identity before we ask, or if they are assigned to the No Label condition.

> Thank you for providing your perspective on the policy proposals. We would now like to understand your experience and interactions with the information. Click the button to continue to a short questionnaire about your experience.

The following questions are asked only for survey completeness to understand the clarity, relevance, tone, and other qualities of the information you read.

1. (Optional) How, if at all, did the information or its author affect your opinions about the policy?
2. Do you think that this information was written by a human, or do you think that it was generated by an artificial intelligence (AI) model?
   - Definitely human-written
   - Possibly human-written
   - Not sure
   - Possibly AI-generated
   - Definitely AI-generated
3. Did you question whether this information was generated by a human or AI before we specifically asked you this?
   - Yes
   - No
4. Please rate your prior experience with conversational artificial intelligence (AI) models or large language model (LLM) tools. Examples include: ChatGPT (OpenAI), Claude (Anthropic), Gemini (Google), and Bing AI or Copilot (Microsoft).
   - I use conversational AI or LLMs more than once a week.
   - I use conversational AI or LLMs about once a week.
   - I use conversational AI or LLMs about once a month.
   - I use conversational AI or LLMs less than once a month.
   - I never use conversational AI or LLMs.
   - I have never heard of conversational AI or LLMs.

Finally, we ask participants about their demographic information:

1. What is the highest level of school you have completed or the highest degree you have received?
   - No formal schooling
   - Did not receive high school diploma
   - High school graduate
   - GED or equivalent
   - Some college
   - 2-year degree (e.g., associate degree)
   - Bachelor's degree
   - Master's degree
   - Professional or academic doctorate degree
   - Prefer not to answer
2. Please choose whichever race and/or ethnicity that you identify with (you may choose more than one option):
   - American Indian or Alaska Native
   - Asian
   - Black or African American
   - Hispanic or Latino
   - Middle Eastern or North African
   - Native Hawaiian or Pacific Islander
   - White
   - Other
3. Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or other?
   - Republican
   - Democrat
   - Independent
   - I don't identify with any political party
   - Other
   - Prefer not to answer

Participants' age is provided directly by Prolific.

C Participants

We required participants to be English-speaking, U.S. residents, with an approval rate of 97-100 and at least 100 prior submissions on Prolific. We paid participants at a rate of $15.00 per hour. Participant demographics are shown in Table 2. The number of participants assigned to each condition is shown in Table 3.

Race/Ethnicity	
White	67.6%
Black or African American	13.1%
Asian	7.5%
Hispanic or Latino	4.4%
Political Party	
Democrat	49.1%
Independent	24.4%
Republican	20.1%
Does not identify with any political party	4.8%
Education	
Bachelor's degree or more	57.9%
Some college	30.0%
High school graduate	10.3%
GED or equivalent	1.1%
Did not receive high school diploma	0.3%
AI Experience	
I use conversational AI or LLMs more than once a week.	19.1%
I use conversational AI or LLMs about once a week.	19.7%
I use conversational AI or LLMs about once a month.	15.4%
I use conversational AI or LLMs less than once a month.	27.2%
I never use conversational AI or LLMs.	17.2%
I have never heard of conversational AI or LLMs.	1.3%

Table 2: Participant demographics.

	AI Label	**Human Label**	**No Label**	**Total**
Geoengineering	127	118	132	377
Drug Importation	128	117	136	381
College Athlete Salaries	126	124	130	380
Social Media Platform Liability	124	119	134	377
Total	505	478	532	1,515

Table 3: Number of participants assigned to each of the policy and authorship label conditions.

D Extended Results

The code repository used to generate all results is available here: `https://github.com/i-gallegos/ai-authorship-persuasion`. We present descriptive statistics in Table 4, and show the change in confidence, sharing intentions, and accuracy judgments across each condition in Figures 7–9. In particular, participants in the AI Label condition increased their support for the policy on average by 9.33 points on the 0-100 scale (SE=0.77), while participants in the Human Label condition increased their support on average by 11.34 points (SE=0.77), with a 9.35 point average increase (SE=0.66) in the No Label condition.

Regression results for the support, confidence, sharing intention, and accuracy judgment are shown in Tables 5–9. While Table 5 shows our main regression model, which models all policies to improve causal inference compared to a single policy, we also show single-policy regression results in Table 6, where the difference between the AI Label and Human Label conditions remains insignificant. Each of the regression results shows that we do not observe a significant effect between the AI Label

	Change in Support	Change in Confidence	Sharing	Accuracy
AI Label	9.33 ± 0.77	7.83 ± 0.79	43.26 ± 1.47	63.46 ± 0.87
Human Label	11.34 ± 0.77	8.55 ± 0.81	45.60 ± 1.48	65.37 ± 0.86
No Label	9.35 ± 0.66	6.79 ± 0.72	45.61 ± 1.35	63.86 ± 0.86

Table 4: Descriptive statistics, with mean and standard error. Change in support and change in confidence represent the difference between the post-intervention and pre-intervention variables. Sharing and accuracy are measured only in the post-intervention stage. Participants in the Human Label condition tend to have the highest change in support, change in confidence, sharing intentions, and accuracy judgments; however, regression results show that none of these differences are significant.

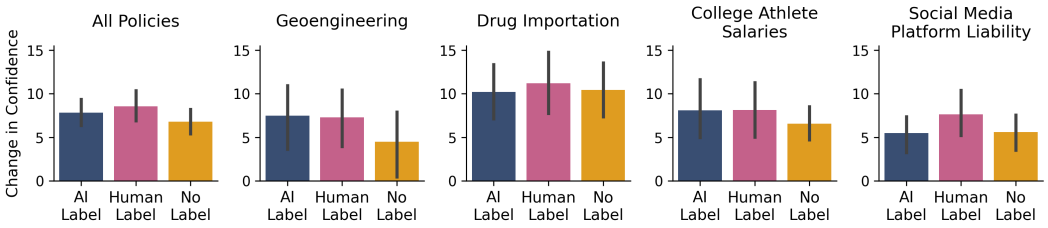

Figure 7: Mean and 95% confidence intervals for the difference between the post-intervention and pre-intervention confidence variables for the AI Label, Human Label, and No Label conditions.

condition and the Human Label condition or the No Label condition. The equivalence bounds for the support variable from the TOST equivalence tests are shown in Figure 10.

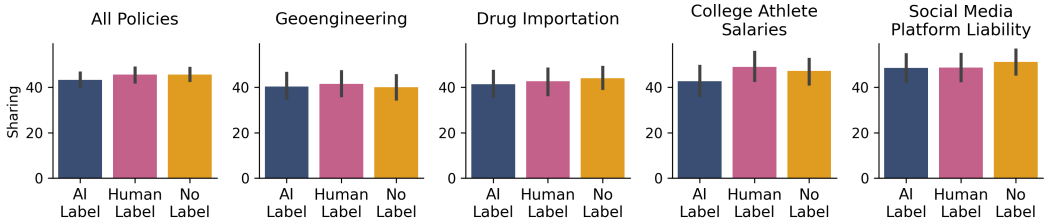

Figure 8: Mean and 95% confidence intervals for the post-intervention sharing intention variable for the AI Label, Human Label, and No Label conditions.

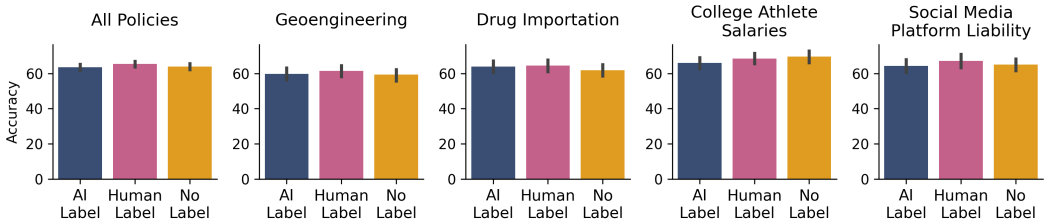

Figure 9: Mean and 95% confidence intervals for the post-intervention accuracy judgment variable for the AI Label, Human Label, and No Label conditions.

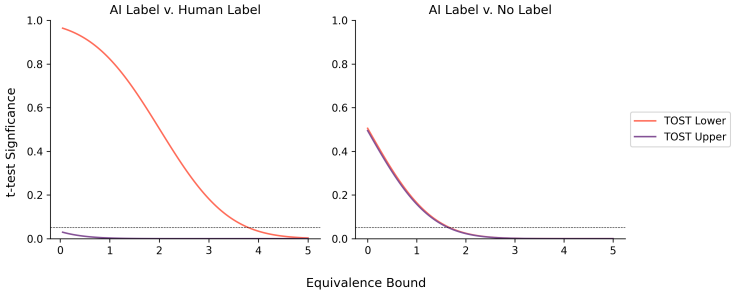

Figure 10: Statistical significance of equivalence versus equivalence bounds for the difference between the post-intervention and pre-intervention support variables. For lower and upper equivalence limits EL and EU, TOST Lower represents the t-test for the hypothesis $H_0 : \mu_i - \mu_{\text{AI Label}} \leq EL$, while TOST Upper represents the t-test for the hypothesis $H_0 : \mu_i - \mu_{\text{AI Label}} \geq EU$.

Model	Variable	b	S.E.	p	95% CI Lower	95% CI Upper
Model 1	(Intercept)	25.49***	1.23	<0.001	23.07	27.90
	Pre-Intervention DV	0.83***	0.01	<0.001	0.80	0.86
	AI Label	−1.82 †	0.98	0.063	-3.74	0.10
	No Label	−1.73 †	0.97	0.074	-3.63	0.17
	R^2=0.674					
Model 2	(Intercept)	23.76***	1.22	<0.001	21.35	26.16
	Pre-Intervention DV	0.83***	0.01	<0.001	0.80	0.86
	AI Label	−0.09	0.95	0.926	-1.96	1.78
	Human Label	1.73†	0.97	0.074	-0.17	3.63
	R^2=0.674					
Moderator: Party	(Intercept)	24.00***	3.42	<0.001	17.29	30.71
	Pre-Intervention DV	0.82***	0.01	<0.001	0.80	0.85
	AI Label	−2.42	4.20	0.565	-10.66	5.82
	No Label	0.53	4.08	0.896	-7.46	8.52
	Democrat x AI Label	−0.35	4.44	0.938	-9.06	8.37
	Republican x AI Label	4.10	4.71	0.384	-5.14	13.35
	Independent x AI Label	−0.40	4.63	0.930	-9.48	8.67
	Democrat x No Label	−2.24	4.31	0.603	-10.68	6.21
	Republican x No Label	−3.06	4.63	0.508	-12.14	6.01
	Independent x No Label	−2.79	4.51	0.537	-11.64	6.07
	R^2=0.677					
Moderator: Prior Knowledge	(Intercept)	25.40***	1.37	<0.001	22.72	28.09
	Pre-Intervention DV	0.83***	0.02	<0.001	0.81	0.86
	AI Label	−0.76	1.35	0.576	-3.41	1.90
	No Label	−1.38	1.35	0.307	-4.03	1.27
	Knowledge x AI Label	−4.92	4.28	0.250	-13.32	3.47
	Knowledge x No Label	−1.57	4.26	0.712	-9.93	6.78
	R^2=0.677					
Moderator: AI Experience	(Intercept)	26.16***	1.85	<0.001	22.53	29.79
	Pre-Intervention DV	0.83***	0.01	<0.001	0.80	0.86
	AI Label	−2.87	2.24	0.201	-7.27	1.53
	No Label	−1.98	2.19	0.365	-6.28	2.31
	Experience x AI Label	1.80	3.44	0.601	-4.95	8.55
	Experience x No Label	0.45	3.40	0.895	-6.22	7.11
	R^2=0.675					
Moderator: Education	(Intercept)	25.02***	2.20	<0.001	20.70	29.35
	Pre-Intervention DV	0.83***	0.01	<0.001	0.80	0.86
	AI Label	2.18	2.80	0.435	-3.30	7.67
	No Label	−1.98	2.81	0.482	-7.49	3.54
	Education x AI Label	−6.71	4.37	0.125	-15.29	1.87
	Education x No Label	0.46	4.31	0.915	-7.98	8.91
	R^2=0.676					
Moderator: Age	(Intercept)	27.69***	2.63	<0.001	22.54	32.84
	Pre-Intervention DV	0.83***	0.01	<0.001	0.80	0.86
	AI Label	1.32	3.36	0.695	-5.27	7.90
	No Label	−2.25	3.34	0.501	-8.79	4.30
	Age x AI Label	−0.08	0.08	0.337	-0.24	0.08
	Age x No Label	0.01	0.08	0.866	-0.14	0.17
	R^2=0.676					

Table 5: Regression coefficients for the support variable. Model 1 treats the Human Label condition as the reference category, given by the regression equation Post-Intervention Support $= \beta_0 + \beta_1$(Pre-Intervention Support) $+ \beta_2$(AI Label Condition) $+ \beta_3$(No Label Condition) $+ \beta_4$(Topic 2) $+ \beta_5$(Topic 3) $+ \beta_6$(Topic 4). Model 2 treats the No Label condition as the reference category, given by the regression equation Post-Intervention Support $= \beta_0 + \beta_1$(Pre-Intervention Support) $+ \beta_2$(AI Label Condition) $+ \beta_3$(Human Label Condition) $+ \beta_4$(Topic 2) $+ \beta_5$(Topic 3) $+ \beta_6$(Topic 4). Moderator effects add interaction terms to Model 1. ***p<0.001, **p<0.01, †p<0.10.

Model	Variable	b	S.E.	p	95% CI Lower	95% CI Upper
Geoengineering	(Intercept)	32.80***	2.25	<0.001	28.37	37.24
	Pre-Intervention DV	0.67***	0.03	<0.001	0.61	0.74
	AI Label	−0.22	2.00	0.912	-4.14	3.70
	No Label	−1.39	1.98	0.483	-5.29	2.51
	R^2=0.528					
Drug Importation	(Intercept)	22.97***	2.36	<0.001	18.32	27.61
	Pre-Intervention DV	0.80***	0.03	<0.001	0.73	0.87
	AI Label	−0.81	1.86	0.664	-4.46	2.84
	No Label	−2.58	1.83	0.160	-6.18	1.02
	R^2=0.596					
College Athlete Salaries	(Intercept)	20.53***	2.12	<0.001	16.37	24.70
	Pre-Intervention DV	0.85***	0.03	<0.001	0.79	0.90
	AI Label	−3.52 †	2.09	0.093	-7.62	0.59
	No Label	−4.25 *	2.08	0.042	-8.34	-0.16
	R^2=0.703					
Social Media Platform Liability	(Intercept)	9.26***	1.88	<0.001	5.57	12.95
	Pre-Intervention DV	0.94***	0.02	<0.001	0.89	0.99
	AI Label	−3.08 †	1.78	0.084	-6.59	0.42
	No Label	0.47	1.75	0.786	-2.96	3.91
	R^2=0.795					

Table 6: Regression coefficients for the support variable, where each model considers only a single policy. Each model treats the Human Label condition as the reference category, given by the regression equation Post-Intervention Support $= \beta_0 + \beta_1$(Pre-Intervention Support)$+ \beta_2$(AI Label Condition)$+$ β_3(No Label Condition). ***p<0.001, **p<0.01, †p<0.10.

Model	Variable	b	S.E.	p	95% CI Lower	95% CI Upper
Model 1	(Intercept)	27.19***	1.28	<0.001	24.68	29.69
	Pre-Intervention DV	0.66***	0.01	<0.001	0.63	0.69
	AI Label	−1.09	0.95	0.247	-2.95	0.76
	No Label	−1.32	0.93	0.159	-3.15	0.51
	R^2=0.595					
Model 2	(Intercept)	25.87***	1.27	<0.001	23.38	28.36
	Pre-Intervention DV	0.66***	0.01	<0.001	0.63	0.69
	AI Label	0.22	0.92	0.809	-1.58	2.03
	Human Label	1.32	0.93	0.159	-0.51	3.15
	R^2=0.595					

Table 7: Regression coefficients for the confidence variable. Model 1 treats the Human Label condition as the reference category, given by the regression equation Post-Intervention Confidence $= \beta_0 + \beta_1$(Pre-Intervention Confidence) $+ \beta_2$(AI Label Condition) $+ \beta_3$(No Label Condition) $+ \beta_4$(Topic 2) $+ \beta_5$(Topic 3) $+ \beta_6$(Topic 4). Model 2 treats the No Label condition as the reference category, given by the regression equation Post-Intervention Confidence $= \beta_0 + \beta_1$(Pre-Intervention Confidence) $+ \beta_2$(AI Label Condition) $+ \beta_3$(Human Label Condition) $+ \beta_4$(Topic 2) $+ \beta_5$(Topic 3) $+ \beta_6$(Topic 4). ***p<0.001, **p<0.01, †p<0.10.

Model	Variable	b	S.E.	p	95% CI Lower	95% CI Upper
Model 1	(Intercept)	41.38***	2.04	<0.001	37.37	45.39
	AI Label	−2.27	2.04	0.266	-6.27	1.73
	No Label	0.05	2.01	0.980	-3.90	4.00
	R^2=0.012					
Model 2	(Intercept)	41.43***	1.99	<0.001	37.53	45.33
	AI Label	−2.32	1.99	0.242	-6.22	1.57
	Human Label	−0.05	2.01	0.980	-4.00	3.90
	R^2=0.012					

Table 8: Regression coefficients for the sharing intention variable. Model 1 treats the Human Label condition as the reference category, given by the regression equation Post-Intervention Sharing $= \beta_0 + \beta_1$(AI Label Condition)$+ \beta_2$(No Label Condition)$+ \beta_3$(Topic 2)$+ \beta_4$(Topic 3)$+ \beta_5$(Topic 4). Model 2 treats the No Label condition as the reference category, given by the regression equation Post-Intervention Sharing $= \beta_0 + \beta_1$(AI Label Condition) $+ \beta_2$(Human Label Condition) $+$ β_3(Topic 2) $+ \beta_4$(Topic 3) $+ \beta_5$(Topic 4). ***p<0.001, **p<0.01, †p<0.10.

Model	Variable	b	S.E.	p	95% CI Lower	95% CI Upper
Model 1	(Intercept)	61.16^{***}	1.23	<0.001	58.75	63.58
	AI Label	-1.84	1.23	0.134	-4.25	0.57
	No Label	-1.44	1.21	0.235	-3.82	0.94
	R^2=0.024					
Model 2	(Intercept)	59.73^{***}	1.20	<0.001	57.38	62.07
	AI Label	-0.40	1.19	0.736	-2.75	1.94
	Human Label	1.44	1.21	0.235	-0.94	3.82
	R^2=0.024					

Table 9: Regression coefficients for the accuracy judgment variable. Model 1 treats the Human Label condition as the reference category, given by the regression equation Post-Intervention Accuracy $=$ $\beta_0 + \beta_1$(AI Label Condition) $+ \beta_2$(No Label Condition) $+ \beta_3$(Topic 2) $+ \beta_4$(Topic 3) $+ \beta_5$(Topic 4). Model 2 treats the No Label condition as the reference category, given by the regression equation Post-Intervention Accuracy $= \beta_0 + \beta_1$(AI Label Condition) $+ \beta_2$(Human Label Condition) $+$ β_3(Topic 2) $+ \beta_4$(Topic 3) $+ \beta_5$(Topic 4). ***p<0.001, **p<0.01, †p<0.10.