US Centre Summer Research Grant

Recipient name: Isolde Hegemann

Project title: The effects of different fact-checking interventions on engagement with political misinformation among Republicans

Summary of project:

What fact-checking interventions are most effective at curbing engagement with social media posts that contain misleading political statements? To answer this question, I recruited 1450 US respondents who identify as Republicans on Prolific and set an experimental vignette design in a simulated social media environment. Throughout the visual survey experiment, participants could interact with posts containing true and false political statements like they would in real life. The data collection took place between September 8th and 16th 2025. I find that fact-checking generally works at increasing discernment, both by reducing engagement with posts containing false political statements and increasing engagement with those containing true statements. AI and Community Notes are significantly more effective at reducing overall engagement with misleading posts than independent fact-checkers. AI also remains significantly more effective than independent fact- checkers in a number of robustness checks and exploratory analyses. These findings have policy-relevant and practical implications for how misleading content can be moderated online to avoid its spread.

This work was supported by the LSE Phelan United States Centre and the Economic and Social Research Council.

Introduction

In this study, I investigate how different fact-checking strategies - independent fact-checkers, Community Notes, and AI - influence how users engage with political misinformation on social media. For this purpose, I focus on Republican users and how they engage with fictional co-partisan posts featuring real statements by Donald Trump. Republican users and pro-Trump misinformation are highly relevant in this context as there is evidence of pro-Trump and conservative X (formerly Twitter) users being much more likely to share links to low-quality sources (Mosleh et al., 2024).

One of the key things we know about fact-checking is that it generally works in correcting people's beliefs in a range of misinformation (Porter & Wood, 2021; Hoes et al., 2024), including political misinformation (Walter et al., 2019). And even if people do not trust fact-checkers, fact-checking labels are still successful at reducing beliefs and sharing of posts labelled as containing misinformation (Martel & Rand, 2024). Therefore, I expect to find evidence for H1:

H1: Fact-checking labels are associated with reduced engagement with posts containing misleading political statements.

However, to what degree people respond to a fact-checking message is dependent on whether people perceive the fact-checker to be credible (Li & Chang, 2022; Liu et al., 2023; Bruns et al., 2024). This suggest that credibility is key when it comes to the efficacy of a fact-checking strategy. While Liu et al. (2023) find no statically significant difference between fact-checking organisations, crowdsourcing, and AI among the general US population (p. 14), 70% of Republicans in 2019 said that fact-checkers 'tend to favour one side' (Walker & Gottfried, 2019). Due to asymmetric patterns of misinformation sharing of different political camps, independent fact-checkers are much more likely to flag misinformation shared by pro-Trump and conservative Twitter accounts, making them the subject of intense criticism (Mosleh et al., 2024). In contrast to independent fact-checkers, I expect Community Notes as less likely to be perceived as partisan by Republicans as they require users who hold a range of political beliefs to agree on flagging a post as containing misinformation:

H2: Community notes are more effective than labels by independent fact-checkers in decreasing engagement with posts containing misleading political statements.

When comparing the efficacy of independent fact-checkers and AI among Republicans, Yaqub et al. (2020) find that independent fact-checkers are the most effective at reducing sharing intent, while AI even increases sharing intent. However, as AI performance has significantly improved and become an integral part of everyday experiences since 2020, these findings should be revisited. In the analysis, I expect to find evidence for the following hypotheses:

H3: AI-powered notes are less effective than labels by independent fact-checkers in decreasing engagement with posts containing misleading political statements.

H4: AI-powered notes are less effective than community notes in decreasing engagement with posts containing misleading political statements.

Fact-checking labels only work if people engage with their content and recognise whether a post is labelled as true or false. Therefore, I also examine H5:

H5: Fact-checking labels are associated with increased discernment in engagement between posts containing true versus misleading political statements.

Empirical Strategy

Research Design

The study was pre-registered on OSF and executed in line with the pre-registration.[1] The data collection ran from 8th to 16th of September 2025. Participants were recruited on the platform Prolific and their self-declared political affiliation had to be Republican. They also had to have provided information for basic demographic screeners and pass an attention check to proceed to the experimental part of the survey. My target sample size was 2000 and I had estimated a median completion time of 6 minutes and compensation of £0.75 based on previous studies. However, participants ended up taking a median of 11 minutes which reduced my sample size to 1450 respondents (including platform costs).[2] The sample characteristics and how they compare to population benchmarks of Republicans are summarised in Table 1.

[1] The pre-registration is available under https://osf.io/8nfd3.

[2] This project was approved by the LSE Ethics Review Board. Participants were informed of the purpose of the research, could withdraw their consent at any point during the survey, and were thoroughly debriefed at the end of the survey.

Table 1. Descriptive sample statistics (n=1450) and population benchmarks of Republicans

Category	Group	Sample n	Sample %	Population %
Ethnicity	White	1207	83.2	79
	Ethnic minority	243	16.8	21
Age	18-29	243	16.8	8
	30-49	719	49.6	27
	50+	488	33.7	65
Sex	Female	911	62.8	47
	Male	539	37.2	53
Education	>= Bachelor	767	52.9	36
	< Bachelor	683	47.1	64

Note: 'Ethnic minority' includes respondents who answered 'Asian', 'Black', 'Mixed', and 'Other'. Percentages rounded to one decimal place. Republican population benchmarks are based on figures by Pew Research Center (2024).

Departing from traditional survey experiments used in similar studies, I created a realistic Facebook-like environment to enable users to interact with posts.[3] This approach has the advantage that it allows participants to engage with posts like they would in the real world, offering a high degree of ecological validity. I used a within- and between-subjects experimental vignette design. Respondents were presented with 17 single profiles to which they could respond. Out of these 17, 12 were posts including misleading political statements. A label was randomly assigned according to the following ratios: Fact-checkers (0.25); Other users (0.25); AI (0.25); with the remaining misleading posts not receiving a label (0.25). The remaining five posts were true political statements, which randomly had either no label assigned (0.4), or had a label attached featuring Fact-checkers (0.2), Other users (0.2), or AI (0.2). Post content, profile names and profile pictures, time of post, number of reactions, and number of reposts were randomly assigned. Examples of how these posts and fact-checking labels appeared to participants can be found in the Appendix.

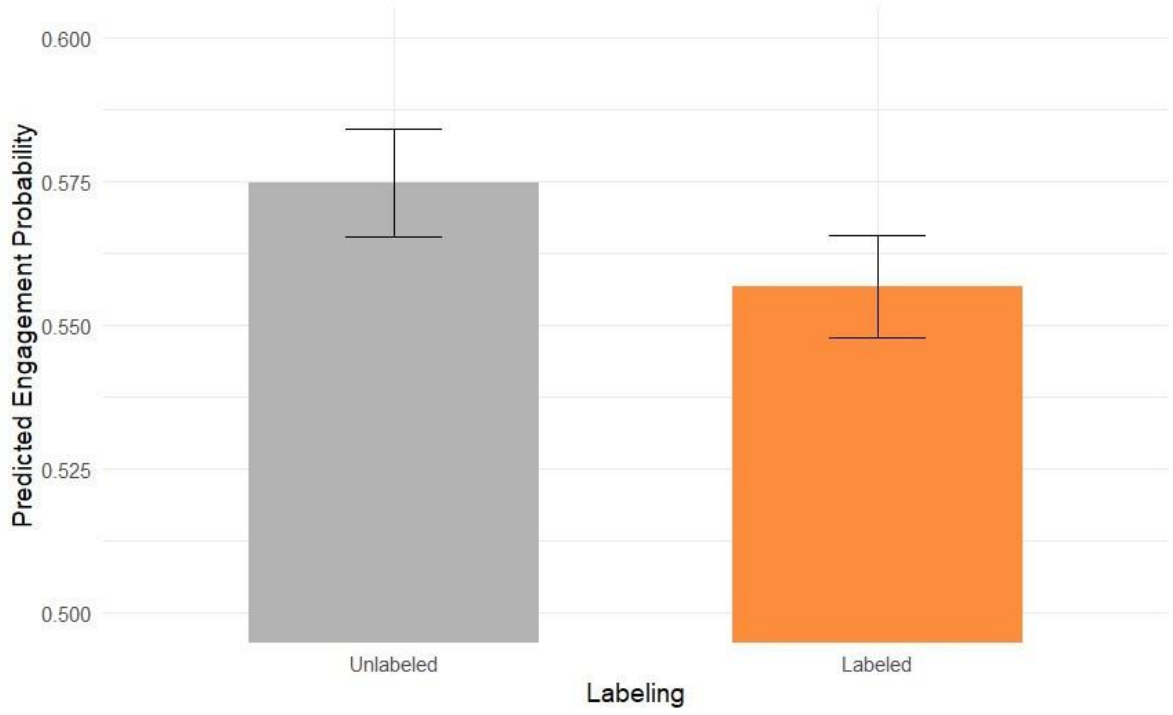
[3] I designed the experiment using the experiment builder 'Gorilla' (www.gorilla.sc).

Analysis and Results

I use a linear OLS model to estimate the effects on engagement probability. Engagement includes liking or sharing a post or commenting on a post. If a participant did not engage with any of the posts, that subject was not included in the main analysis. For all models, standard errors are clustered at the level of the respondent. I use 95% confidence intervals and outliers are included in the analysis.

To examine H1, I regress any engagement with a false post on whether a fact-checking strategy was used or not. As visualised in Figure 1, fact-checking significantly reduces overall engagement with false posts by 1.8 percentage points from 57.5% engagement with unlabelled posts ($p<0.001$).

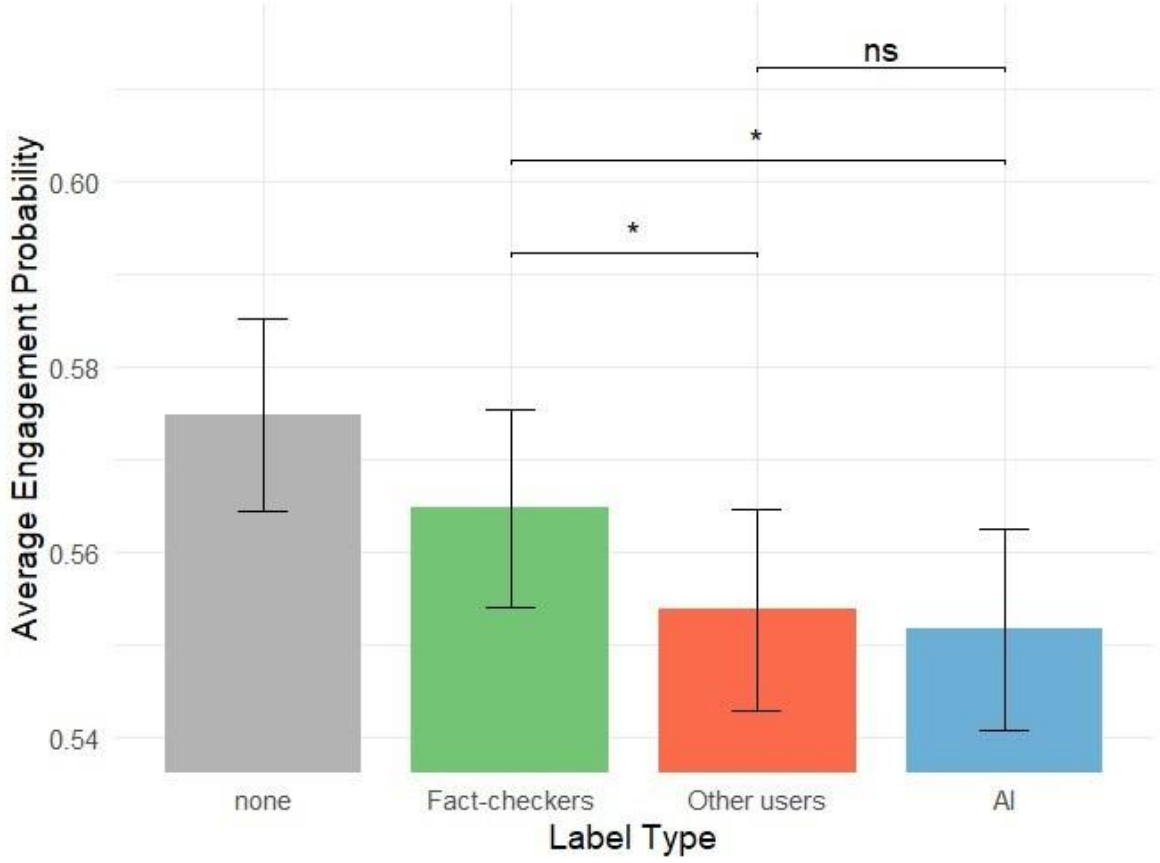
Figure 1. Fact-checking labels decrease engagement with false posts



Note: N=1209 respondents. Includes outliers, excludes non-engagers. Standard errors are clustered at the respondent level. Error bars represent 95% CIs. See Appendix for corresponding table (Table A2).

Turning to H2 to H4, I examine whether some fact-checking strategies are more effective than others in decreasing engagement with false posts. I regress any type of engagement with a post on the different fact-checking strategies and use linear contrasts to examine whether any differences in effectiveness of fact-checking strategies are statistically significant. The results are visualised in Figure 2: In line with H2, other users are more effective at curbing engagement with false information compared to fact-checkers, with a difference of 1.1 percentage points ($p<0.05$). However, I cannot reject the null hypotheses for H3 and H4: Against my expectations, AI is more effective than fact-checkers at decreasing engagement with false posts, with a difference of 1.3 percentage points ($p<0.05$). Furthermore, there is no statistically significant difference in the effectiveness of AI and Community Notes.

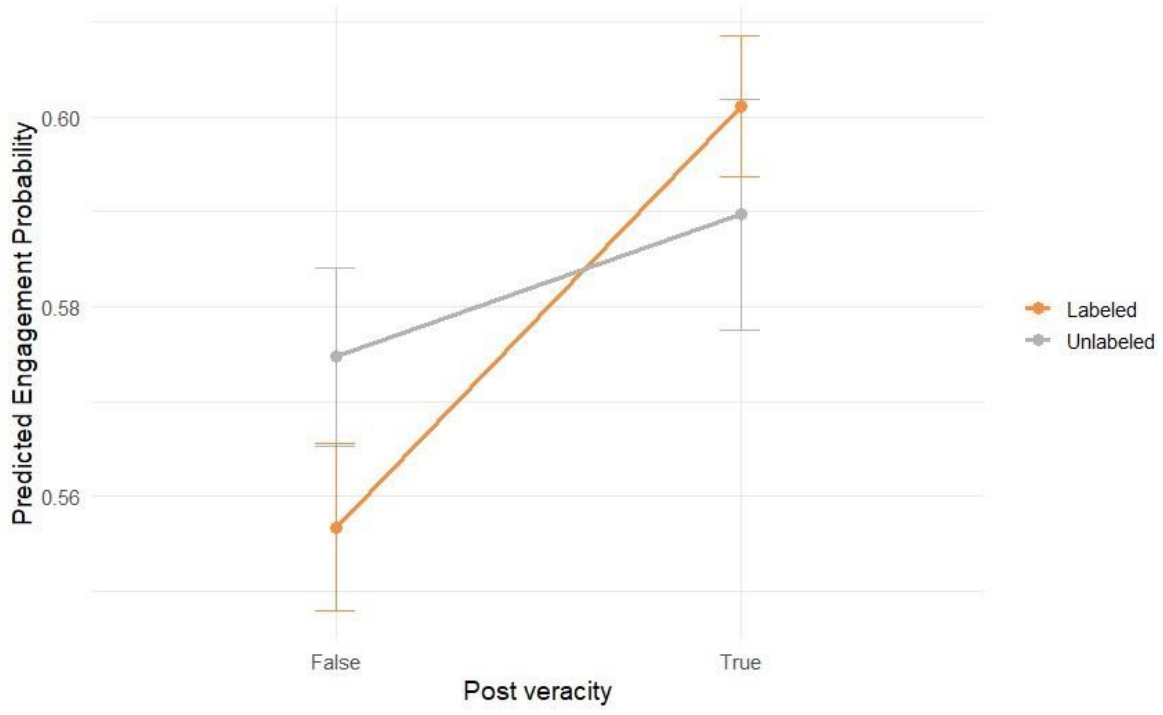
Figure 2. AI and Other users are more effective than Fact-checkers at decreasing engagement



Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. N=1209 respondents. Includes outliers, excludes non-engagers. Standard errors are clustered at the respondent level. Error bars represent 95% CIs. See Appendix for corresponding tables (Table A3 and A4).

To examine H5, I regress engagement with a post on whether a post is true or false, whether that post has a fact-checking label attached to it, and the interaction between both variables. As can be seen in Figure 3, while attaching any fact-checking label to a true statement does not have a statistically significant effect, the interaction between truth value and labelling is statistically significant ($p<0.001$). We can reject the null hypothesis, with the results supporting the expectation that fact-checking labels are associated with increased discernment in engagement between posts containing true versus misleading political statements.

Figure 3. Fact-checking increases discernment between true and false statements



Note: N=1209 respondents. Includes outliers, excludes non-engagers. Standard errors are clustered at the respondent level. Error bars represent 95% CIs. See Appendix for corresponding table (Table A5).

For robustness checks, I run the analysis with all 1450 participants, including those 241 respondents (16.6% of the sample) who do not engage with any post. The patterns from the main analysis remain consistent. I also run the main analysis using weights to make the sample representative of US Republicans by gender, age, ethnicity, and education (Pew Research Center, 2024) and run an exploratory analysis of the main models controlling for the same

covariates. In both the weighted analyses and the covariate-adjusted models, the results largely hold. An exception is the difference between independent fact-checkers and Community Notes which does not reach statistical significance. The main models, robustness checks, and exploratory tests can be found in the Appendix.

Relevance

This study finds that fact-checking generally works in reducing online engagement with co-partisan misinformation among Republicans. However, the type of fact-checking strategy matters: Community Notes and AI are similarly effective at decreasing engagement with false information, while independent fact-checkers do not influence the engagement behaviour of users. Drawing on both the clear statistical significance of the effects in the main analysis and their robustness in further checks, AI-powered fact-checking emerges as an effective strategy to reduce engagement with false co-partisan political statements among US Republicans.

While the effect sizes are modest, they still represent realistic, policy-relevant reductions, with important practical implications for our understanding of the ability of fact-checking strategies to decrease engagement with political misinformation. Further research focused on curbing the online spread of political misinformation in the US should explore the effects of Community Notes and AI among different groups and compare the feasibility and effectiveness of the strategies at different points in the fact-checking process.

Bibliography

Barrera, O., Guriev, S., Henry, E., & Zhuravskaya, E. (2020). Facts, alternative facts, and fact checking in times of post-truth politics. *Journal of Public Economics, 182,* 104123. https://doi.org/10.1016/j.jpubeco.2019.104123

Bruns, H., Dessart, F. J., Krawczyk, M., et al. (2024). Investigating the role of source and source trust in prebunks and debunks of misinformation in online experiments across four EU countries. *Scientific Reports, 14,* 20723. https://doi.org/10.1038/s41598-024-71599-6

Hoes, E., Aitken, B., Zhang, J., et al. (2024). Prominent misinformation interventions reduce misperceptions but increase scepticism. *Nature Human Behaviour, 8*, 1545–1553. https://doi.org/10.1038/s41562-024-01884-x

Li, J., & Chang, X. (2023). Combating misinformation by sharing the truth: A study on the spread of fact-checks on social media. *Information Systems Frontiers, 25,* 1479–1493. https://doi.org/10.1007/s10796-022-10296-z

Liu, X., Qi, L., Wang, L., & Metzger, M. J. (2023). Checking the fact-checkers: The role of source type, perceived credibility, and individual differences in fact-checking effectiveness. *Communication Research.* Advance online publication. https://doi.org/10.1177/00936502231206419

Martel, C., & Rand, D. G. (2024). Fact-checker warning labels are effective even for those who distrust fact-checkers. *Nature Human Behaviour, 8,* 1957–1967. https://doi.org/10.1038/s41562-024-01973-x

Mosleh, M., Yang, Q., Zaman, T., et al. (2024). Differences in misinformation sharing can lead to politically asymmetric sanctions. *Nature, 634,* 609–616. https://doi.org/10.1038/s41586-024-07942-8

Pew Research Center. (2024). *Changing partisan coalitions in a politically divided nation.* https://www.pewresearch.org/politics/2024/04/09/changing-partisan-coalitions-in-a-politically-divided-nation/

Porter, E., & Wood, T. J. (2021). The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proceedings of the National Academy of Sciences, 118(37),* e2104235118. https://doi.org/10.1073/pnas.2104235118

Walker, M., & Gottfriend, J. (2019). Republicans far more likely than Democrats to say fact-checkers tend to favor one side. *Short Read*. Pew Research Center. https://pewrsr.ch/2Fz9e22

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2019). Fact-checking: A meta-analysis of what works and for whom. *Political Communication, 37(3),* 350–375. https://doi.org/10.1080/10584609.2019.1668894

Yaqub, W., Kakhidze, O., Brockman, M. L., Memon, N., & Patil, S. (2020). Effects of credibility indicators on social media news sharing intent. In *Proceedings of the 2020 CHI conference on human factors in computing systems (pp. 1–14)*. Association for Computing Machinery. https://doi.org/10.1145/3313831.3376213

Appendix

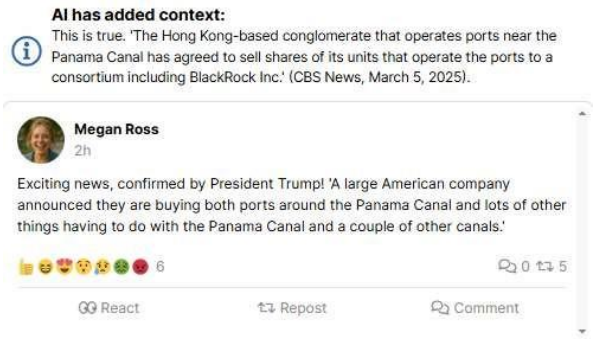
Post examples

Table A1. Visual examples of posts x fact-checking strategies

a) False statement x no context added

b) False statement x fact-checkers

c) False statement x other users



d) True statement x AI

Note: Designed using the experiment builder 'Gorilla' (www.gorilla.sc).

Main models and robustness checks

Table A2. Fact-checking (any)

Variables	(1)	(2)	(3)
Label (any)	-0.018***	-0.023***	-0.017**
	(0.004)	(0.005)	(0.006)
Constant	0.575***	0.533***	0.576***
	(0.005)	(0.006)	(0.006)
Observations	1209	1450	1209
Non-engagers	No	Yes	No
Weights	No	No	Yes

Notes: *** p < 0.001, ** p < 0.01, * p < 0.05. Main model (1), robustness checks (2; 3). Includes outliers. Standard errors (in parentheses) are clustered at the respondent level. Models are covariate-unadjusted (see Exploratory analysis). Observations stand for number of participants. Weights include gender, age, ethnicity, and education (Pew Research Center, 2024).

Table A3. Different fact-checking strategies

Variables	(1)	(2)	(3)
Fact-checkers	-0.010	-0.015*	-0.009
	(0.005)	(0.006)	(0.008)
Other users	-0.021***	-0.027***	-0.018*
	(0.005)	(0.005)	(0.007)
AI	-0.023***	-0.028***	-0.025**
	(0.005)	(0.005)	(0.008)
Constant	0.575***	0.533***	0.576***
	(0.005)	(0.006)	(0.006)
Observations	1209	1450	1209
Non-engagers	No	Yes	No
Weights	No	No	Yes

Notes: *** p < 0.001, ** p < 0.01, * p < 0.05. Main model (1), robustness checks (2; 3). Includes outliers. Standard errors (in parentheses) are clustered at the respondent level. Models are covariate-unadjusted (see Exploratory analysis). Observations stand for number of participants. Weights include gender, age, ethnicity, and education (Pew Research Center, 2024).

Table A4. Linear comparisons between fact-checking strategies

Comparison	(1)	(2)	(3)
Other users vs Fact-checkers	-0.010*	-0.011	-0.009
AI vs Fact-checkers	-0.023*	-0.012*	-0.013*
AI vs Other users	-0.002	-0.002	-0.001
Observations	1209	1450	1209
Non-engagers	No	Yes	No
Weights	No	No	Yes

Note: *** p < 0.001, ** p < 0.01, * p < 0.05. Based on models in Table A3.

Table A5. Fact-checking and discernment

Variables	(1)	(2)	(3)
True post	0.015*	0.015*	0.018*
	(0.006)	(0.007)	(0.007)
Label (any)	-0.018***	-0.023***	-0.017**
	(0.004)	(0.005)	(0.006)
True post x Label (any)	0.029***	0031***	0.030***
	(0.008)	(0.008)	(0.009)
Constant	0.575***	0.533***	0.576***
	(0.005)	(0.006)	(0.006)
Observations	1209	1450	1209
Non-engagers	No	Yes	No
Weights	No	No	Yes

Notes: *** p < 0.001, ** p < 0.01, * p < 0.05. Main model (1), robustness checks (2; 3). Includes outliers. Standard errors (in parentheses) are clustered at the respondent level. Models are covariate-unadjusted (see Exploratory analysis). Observations stand for number of participants. Weights include gender, age, ethnicity, and education (Pew Research Center, 2024).

Exploratory analysis

Table A6. Covariate-adjusted main models

Predictors of interest	Model 1: Any label	Model 2: Label type	Model 3: Discernment
Label (any)	-0.018***		-0.018***
Other users vs Fact-checkers		-0.011	
AI vs Fact-checkers		-0.013*	
AI vs Other users		-0.002	
True post			0.015*
True post x Label (any)			0.030***

Notes: *** p < 0.001, ** p < 0.01, * p < 0.05. Main analysis, covariate-adjusted for exploratory purposes. Includes outlier, excludes non-engagers. Standard errors are clustered at the respondent level. Covariates include gender, age, ethnicity, and education.