Methodological Appendix for

In Disclaimers We Trust: The Effectiveness of State-Required AI Disclaimers on Political Ads

Participants

This study used a convenience sample of N=1051 US adults. Participants were recruited using Prolific, an online survey research platform that provides engaged, high-quality respondents. Participants were eligible for this study if they were at least 18 years old at the time of the survey, a resident of the United States, and indicated that they identified with or leaned toward one of the two major U.S. political parties (i.e., not unaffiliated or independent). After completing the survey, participants received $2.75, or the equivalent of an hourly wage of $16.50. All study procedures were reviewed and designated as exempt from review by the University of North Carolina Institutional Review Board (24-0460).

Procedure

This study used a 3 (candidate party: Democrat, Republican, unaffiliated) x 3 (ad disclaimer: no disclaimer, Michigan label, Florida label) between-persons design with nine conditions. After being directed to a Qualtrics survey, individuals interested in the study provided consent to participate and shared basic demographic information (i.e., age, race, gender) and confirmed their political affiliation. They were then randomized to one of nine experimental conditions and viewed two fictional political advertisements. Each participant viewed one attack ad and one positive ad, and the order of the ads was randomized. After each ad, participants answered a series of questions about their willingness to engage with the ad in various ways (e.g., flag it, dislike it, mute the posting account), perceived accuracy, their impression of the candidate, and whether they would vote for the candidate if they were in the correct district. Last, after viewing both ads, participants were asked about their recall of the disclaimers in the ads as well as their opinion on policies regulating the use and disclosure of generative AI in political advertising.

Stimuli

Stimuli included eighteen political advertisements (nine attack ads and nine positive ads) created specifically for this study. Attack ads featured two fictional candidates for county commissioner in real counties, with a focal candidate accusing an opposition candidate of relying on campaign donations from outside the county. The attack ads featured a series of still images with what was purported to be “real” audio from the opponent saying he would “take money from whoever will give it to me,” and voiceover followed by a short statement by the focal candidate (played by an actor). Positive ads were biographical and featured a fictional candidate for county commissioner (played by an actor) describing his upbringing in the county. The positive ad featured two still images, one showing a young child at the beach, the other a government building.

The party affiliation of candidates featured in the ads varied by condition. One version of each type of ad did not specify party affiliation, one version featured a Democratic candidate

(with Republican opposition in the attack ad), and one featured a Republican candidate (with Democratic opposition in the attack ad). Candidate party was manipulated by including the corresponding party symbol in the intro of the attack ad (elephant/donkey) and by specifying "Democrat" or "Republican" in the outro card for both types of ads. Non-partisan ads did not contain party symbols or text specifying a party.

Additionally, the type of AI label varied by condition. Labels followed design specifications of existing laws in Michigan and Floridia and were included as white text on a black background that appeared at the bottom of the video and remained for the duration of the ad. For each attack and positive ad, three versions were created: One version with no label (control), one with the label text required in Michigan ("This video has been manipulated by technical means and depicts speech or conduct that did not occur."), and one containing the text required in Florida ("This video was created in whole or in part with the use of generative artificial intelligence."). The composition of conditions and links to stimuli are provided in Table 1.

Table 1

Composition of experimental conditions and link to associated stimuli

AI label type	Candidate party		
	Non-partisan	Democrat	Republican
No label	Attack ad Positive ad	Attack ad Positive ad	Attack ad Positive ad
Michigan text	Attack ad Positive ad	Attack ad Positive ad	Attack ad Positive ad
Florida text	Attack ad Positive ad	Attack ad Positive ad	Attack ad Positive ad

Measures

Engagement. We measured participants' willingness to engage with each ad using six items asking about individual actions they might take. Items included typical reactions to social media such as flagging a post, liking/disliking, and commenting. Willingness to perform each action was measured on 5-point scales from 1 ("Not at all willing") to 5 ("Extremely willing").

Candidate perceptions. We assessed participants' perceptions of the candidates in each ad with four items. Items asked about candidate trustworthiness, appeal, intentions to vote for the candidate, and intentions to search for more information about the candidate. Responses were measured on 5-point scales from 1 (e.g., "extremely untrustworthy," "extremely unappealing") to 5 (e.g. "extremely trustworthy," "extremely appealing").

Ad perceptions. We assessed perceptions of each ad with two questions that asked participants the degree to which they felt the ad told the truth and whether it tried to manipulate them. Responses were measured on 5-point scales from 1 ("strongly disagree") to 5 ("strongly

agree"). We also asked participants what types of technical means they believed were used to produce the images in the ad. Participants were asked to select all that apply from a list including "Generative Artificial Intelligence," "Photoshop or photo editing," "Video editing," "Audio editing," and "Paid actors."

Label recall. After participants viewed both ads, we assessed recall of the disclaimers using two items. First, participants were asked directly whether there were disclaimers about generative AI or other technology in the ads they just watched. Responses included "yes," "no," and "not sure." They were then provided the text of three disclaimers (two included in the experiment and one red herring) and asked to select the disclaimer that they saw in the ads they watched. Participants were also provided the option to respond that there were no disclaimers in the ads.

Policy opinions. We assessed participants' attitudes toward the inclusion of disclaimers on political ads using three items. Each item was a statement of a potential policy position.

- "All political ads containing images, video, or audio created through generative AI should include disclaimers."
- "Only political ads containing images, video, or audio created through generative AI that show speech, images, or conduct that did not occur should include disclaimers."
- "Generative AI should be prohibited in political advertisements or campaigns."

Responses were measured on 5-point scales from 1 ("strongly disagree") to 5 ("strongly agree").

Political affiliation and congruence. We measured political affiliation with two items. A first item asked participants whether they consider themselves to be a Republican, Democrat, Independent, or something else. Respondents who answered "Independent" or "Something else" were shown a follow-up question asking whether they leaned more toward the Republican or Democratic party. Respondents who indicated they lean toward a party were coded as a member of that party, and those who indicated they do not lean toward either party were excluded from the study.

We used participants' political affiliation and candidate ad condition (non-partisan, Democrat, or Republican) to create a political congruence variable for use in the moderation analyses. Participants whose ad condition matched their own affiliation were coded as congruent, those who belonged to the opposite party were coded as incongruent, and those who saw non-partisan ads were coded as non-partisan.

Analyses

We analyzed differences in label recall and policy opinions descriptively. To examine differences between label conditions in engagement, candidate perceptions, and ad perceptions, we conducted one-way Analyses of Variance (ANOVA) with type III sum of squares. For the moderation analyses, we conducted two-way ANOVAs and examined the interaction effect of political congruence and label condition. Where interaction effects were significant, we conducted post-hoc pairwise *t*-tests and interpreted results at a Bonferroni-corrected significance level of $p<.017$. For items in which the assumption of

homoskedasticity was violated, heteroskedasticity-robust (Hubert-White) standard errors and unpooled variances were used for tests.