Measuring the Effect of AI Deepfake Videos on Favorabilities of Public Figures

Woojae Chung, Nicole Kan, Joan Lee, Ricky Pang, Mohammad Kanawati

08/14/2024

Abstract

Does fake, AI-generated content on public figures, even if clearly labeled as fake content generated by artificial intelligence, cause people to form more extreme views of the individuals? Realistic content depicting highly influential individuals saying or doing something provocative that they have not said or done can have manipulative and polarizing social consequences. In response, generative AI products have mandated (or at least are considering) including disclaimers that the content may be inaccurate or artificial. However, AI-generated content may still instill bias in the content viewers, and it is unclear whether the inclusion of disclaimers has an effect. Our study intends to experiment to answer this question. By surveying participants on favorability towards various well-known public figures after showing AI-generated videos with and without a disclaimer, and comparing the results to those who weren't shown any videos, we see the effects of deepfake videos on the perception of the public figures. Our experimental study finds that the content generated by AI generally does not affect participants' favorabilities of individuals, but it does have significant results among certain groups. Our findings have important implications for the safe use of AI and the potentially harmful effect it can have on social matters.

Background

The theory behind the experiment is justified by extensive research showing that AI-generated content, including deepfakes, significantly impacts public opinion and trust. Studies indicate that such content spreads rapidly through social media, eroding trust in legitimate information sources and making individuals more susceptible to extreme views. For example, research from Cambridge University highlights how AI facilitates the wide dissemination of disinformation, reaching large audiences quickly (Bontridder & Poullet, 2021). Additionally, MIT research has shown that emotionally charged AI-generated content can deeply influence attitudes, often more than factual corrections, due to cognitive biases (Wittenburg et al., 2023).

The specific focus of this experiment—examining the effect of controversial AI-generated videos on participants' perceptions–is grounded in findings like those of Vaccari and Chadwick (2020), who demonstrated that deepfakes decrease trust in media and increase uncertainty about information authenticity. Given these insights, our study seeks to contribute to understanding the socio-political impacts of AI-generated content and explore interventions, such as labeling content as AI-generated, to mitigate its effects.

Hypothesis

We aim to evaluate the impact of AI-generated deepfake videos on participants' favorability towards familiar figures, among other measures of perception. Based on previous research, we hypothesize that displaying deepfake videos will statistically decrease participants' favorability towards all individuals, and displaying a disclaimer that the content is fake will decrease the magnitude of the decline. This suggests that the realistic and emotionally provocative nature of the content may trigger cognitive and emotional responses, leading to altered perceptions. Formally:

- Null Hypothesis (H0): Viewing AI-generated content will not significantly alter participants' favorability ratings of well-known individuals compared to the control group.
- Alternative Hypothesis (H1): Viewing AI-generated content will significantly alter participants' favorability ratings, whether or not the content includes disclaimers.

Experimental Details

Potential Outcomes and Randomization

Participants enter the experiment by agreeing to partake in a survey. Recruitment is serviced using PureSpectrum's online panel, representative of the United States general population. The survey is created using Qualtrics. In the survey, we employ a randomization process provided by Qualtrics to assign participants into one of three groups: Control, Treatment 1, and Treatment 2. This method eliminates any biases that could arise from non-random assignments and ensures that the treatment effects could be causally attributed to the AI-generated videos rather than any pre-existing differences between participants.

The study's operations are as follows:

- Control: Participants rate public figures on a scale of 0 -10 without any prior exposure to AI-generated content.
- Treatment 1: Participants rate favorability after being exposed to an AI-generated video depicting the figures.
- Treatment 2: Participants rate favorability after being exposed to an AI-generated video that is explicitly labeled as such as a disclaimer.

Participant experience is structured to first determine their familiarity with the public figures in question. They are then randomly assigned to one of the three conditions as described above. Following their assignment, treatment groups are each then shown a video of an individual (with the disclaimer for Treatment 2) and then asked about their favorability of that person on an 11-point scale (0 to 10). This process is repeated for each public figure, where the order of the videos is randomized. Following this process, participants respond to demographic questions, including their gender, age, party affiliation, political ideology, and industry of work.

The primary outcome measured is the change in favorability ratings of the figures before and after the treatments. By comparing the ratings from Control with those of Treatment 1 and Treatment 2, the study assesses the impact of AI-generated content on participants' perceptions. Specifically, we examine if the exposure to AI-generated videos, both labeled and unlabeled, results in a significant difference in how favorably participants view the depicted figures. This comparison allows us to measure the influence of AI-generated content on public perception and whether labeling the content as AI-generated affects this influence.

Covariate Balance Check

As our recruitment process was randomized, our sample is generally balanced across the demographic variables, gender, age group, political party, political ideology, and job industry (binary between tech and other industries/not available), as demonstrated in Figure 1. We include party and ideology to better understand

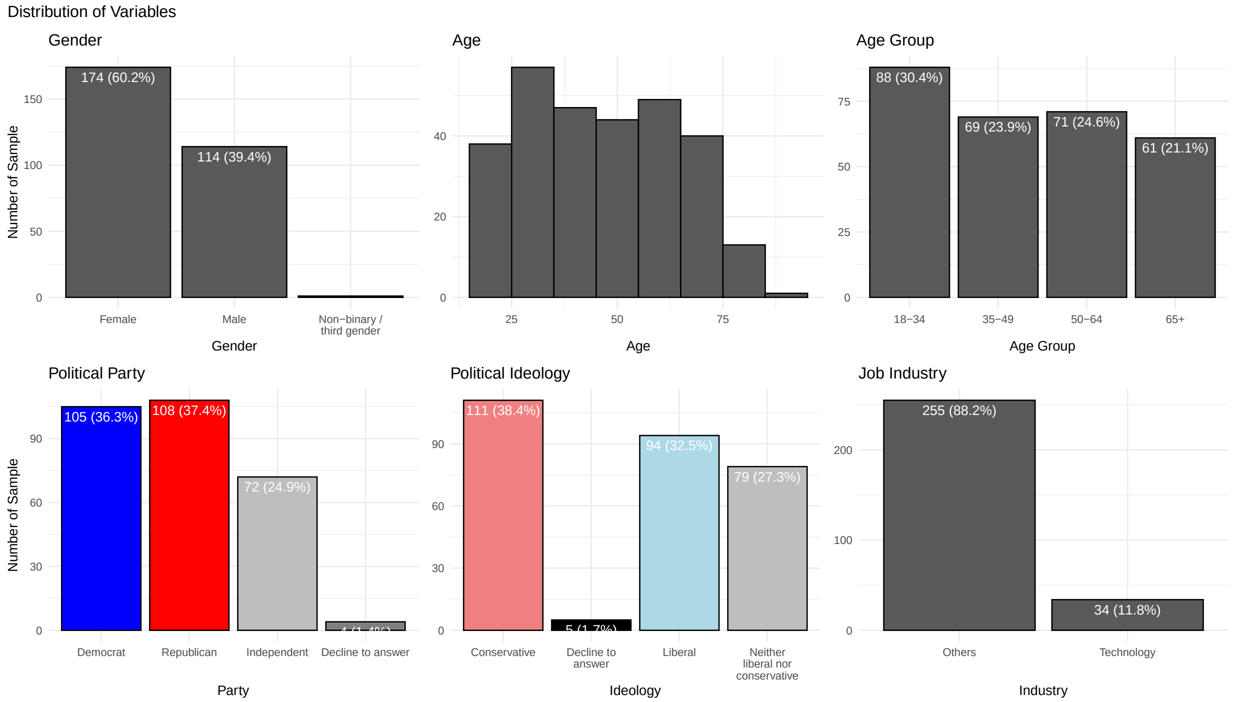

Figure 1: Distribution of Demographic Variables

the treatment effects for certain political groups, given that we believe that one's political beliefs will affect favorability of Biden and Trump, and likewise with the job industry where working in the technology industry may affect favorability for Zuckerberg and Musk.

More importantly, in order to make sound comparisons between the outcome results, the three treatment groups must have comparable samples. We perform a covariate balance check to verify the randomization process by comparing the demographic variables across the three treatment groups. We achieve this with a Chi-square test of independence, which is a common test to determine if there is a meaningful difference between categorical variables. We perform this test on the five categorical demographic variables and examine their p-values. A high p-value (greater than 0.05) suggests that the variable is balanced across the treatment groups, and all of our variables exhibit high p-values. This suggests that our randomization is successful.

Treatment

The treatment in our study is the AI-generated videos of four well-known figures (Joe Biden, Donald Trump, Mark Zuckerberg, Elon Musk). Each video consists of a close-up video showing the individuals' faces as they read a script that we write. The script is created to give a negative realistic impression of the person. We incorporate recent issues from the news in an attempt to make the content match what they would say in reality, and we also match the tone and vocabulary in order to closely follow their true persona. For example, we include stuttering in Joe Biden's script to incorporate the negative media about Biden's stuttering. For treatment 2, we clearly state the disclaimer *"The video you are seeing is generated by artificial intelligence and is not reflective of actual video footage or audio recording"* underneath the videos.

Consort Diagram

PureSpectrum distributed the survey to an online panel of the United States general population in which 419 subjects completed the survey. Out of the 419, 289 met the criteria that they were familiar with all the famous figures in our treatment. Our final sample size of 289 people is randomly assigned to control or treatment and their data is used in our analysis.

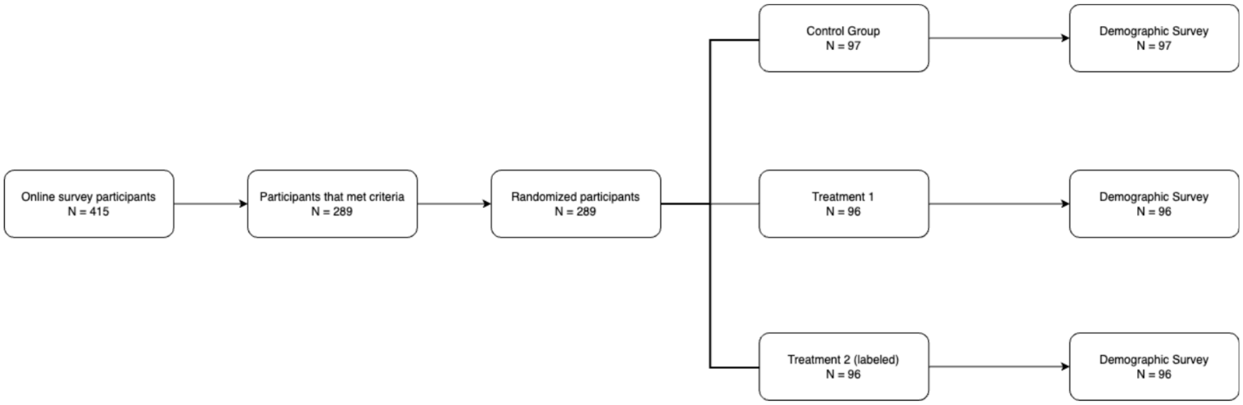

Figure 2: Consort Diagram

Power Calculation

Prior to analyzing the data we obtain, we simulate data to assess the power of detecting different effect sizes (0.2, 0.3, 0.4) between control and treatment groups. We generate normally distributed favorability scores for each group and adjust the means to reflect the effect sizes. The simulation includes varied sample sizes, calculated as percentages of the total data. We conduct t-tests comparing each treatment group against the control group. In over 1,000 simulations, we determine the proportion of tests rejecting the null hypothesis for each sample size, representing the power of the experiment. The results indicate that for Treatment 2, detecting a small effect size of 0.2 required approximately 50 participants per group to achieve a power of 0.80, with power increasing to 100% at 200 participants. For moderate and large effect sizes (0.3 and 0.4), the necessary sample sizes to achieve high power are significantly lower, with about 20 participants per group achieving high power. Repeating for Treatment 1 shows a higher power even with smaller sample sizes across all effect sizes, achieving a power close to 0.80 with just 20 participants for a 0.2 effect size and reaching 100% power with as few as 50 participants. This analysis demonstrates that Treatment 1 requires fewer participants to achieve a high power compared to Treatment 2, especially for detecting small effects, due to the stronger assumed effect of the treatment without a disclaimer.

Analysis

Data

The outcome dependent variable for this experiment, favorability, is measured for each observation for each public figure tested. These favorability scores are examined both on the individual level per public figure and averaged together per observation.

In Figure 3, we see the distribution of favorability scores for Biden, Trump, Zuckerberg, and Musk (rightmost four), as well as the distribution of favorabilities of the four individuals averaged together (leftmost). We observe that while on average, favorability is normally distributed, the underlying distributions are not, especially for Biden and Trump. This is more salient when looking specifically at how the distributions are laid out for Democrats and Republicans, as in Figure 4.

Our survey participants portray extreme views on political candidates, with Democrats showing extremely right-skewed distribution for Trump favorability as the majority rate Trump a zero rating. The same is true

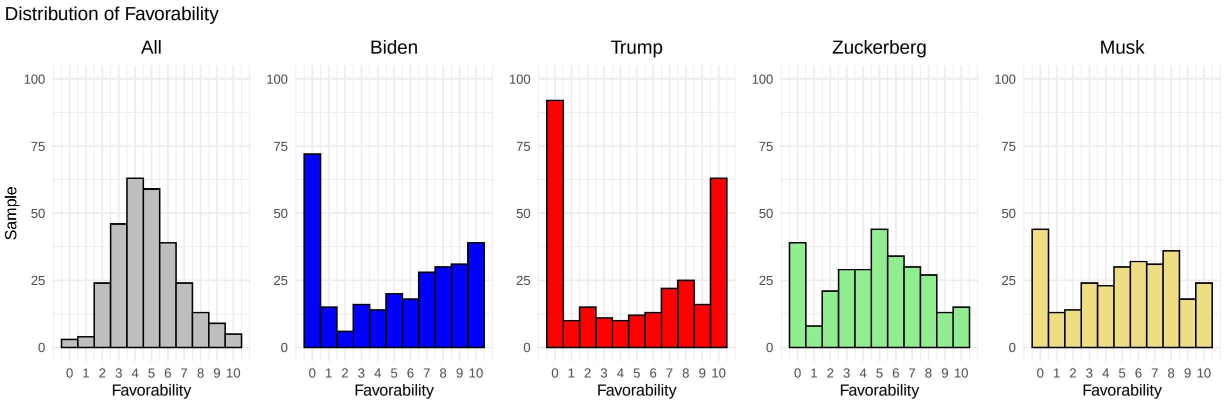

Figure 3: Distribution of Favorability

with Biden for Republicans. On the other hand, while Republicans depict an extremely high favorability of Trump, Democrats' high favorability of Biden is not on the same magnitude. This difference can be a concern for bias when individual favorabilities are considered, as different groups of participants exhibit different favorability measures for each individual tested. For example, a Republican who is assigned to treatment may see a Trump video, but since they are more likely to be blinded by their extreme favorability towards Trump, the video may not be effective for them. Meanwhile, a Democrat who sees a Biden video may feel disappointed by Biden's fake comment because they are less likely to show extreme favorability towards Biden as the Republican does towards Trump. While this is an area that warrants further study, we do not explore this further here. In a similar manner, we explore the possibility of skewed distributions for Zuckerberg and Musk for those in technology job industries, but we do not find noteworthy outcomes.

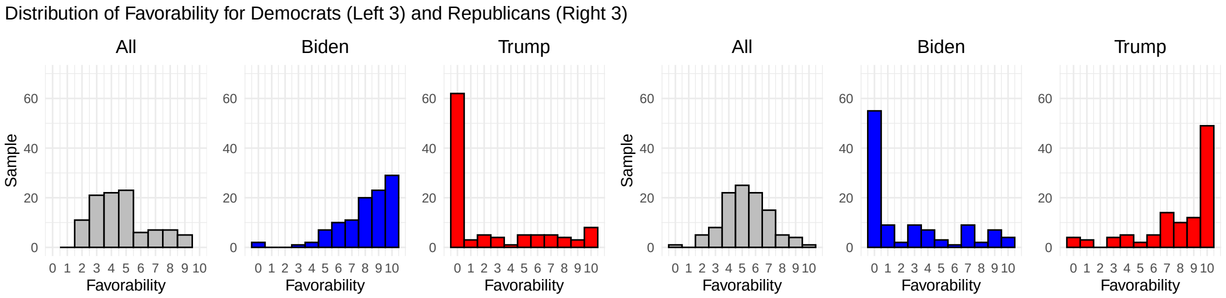

Figure 4: Distribution of Favorability for Democrats (Left) and Republicans (Right)

The most important predictor variable is the treatment assignment, whether the participant is exposed to the deepfake videos showing the public figures making negative comments, and whether the disclaimers indicating that they are fake are present. Favorability distributions are also examined across the three treatment groups. Figure 5 shows the distributions along with the average lines. We do not observe notable differences in the distributions across the three treatment groups. The important covariates that we analyze in the paper are the demographics they belong to, which allow us to sub-group the data and measure how the treatments affect different participants in the study.

Models

While we see that the distributions of favorability scores are similar across the treatment groups for all public figures and the averages, we build linear models to measure observed effects and validate their significance. By utilizing a straightforward linear model to assess the relationship between the treatment variables and

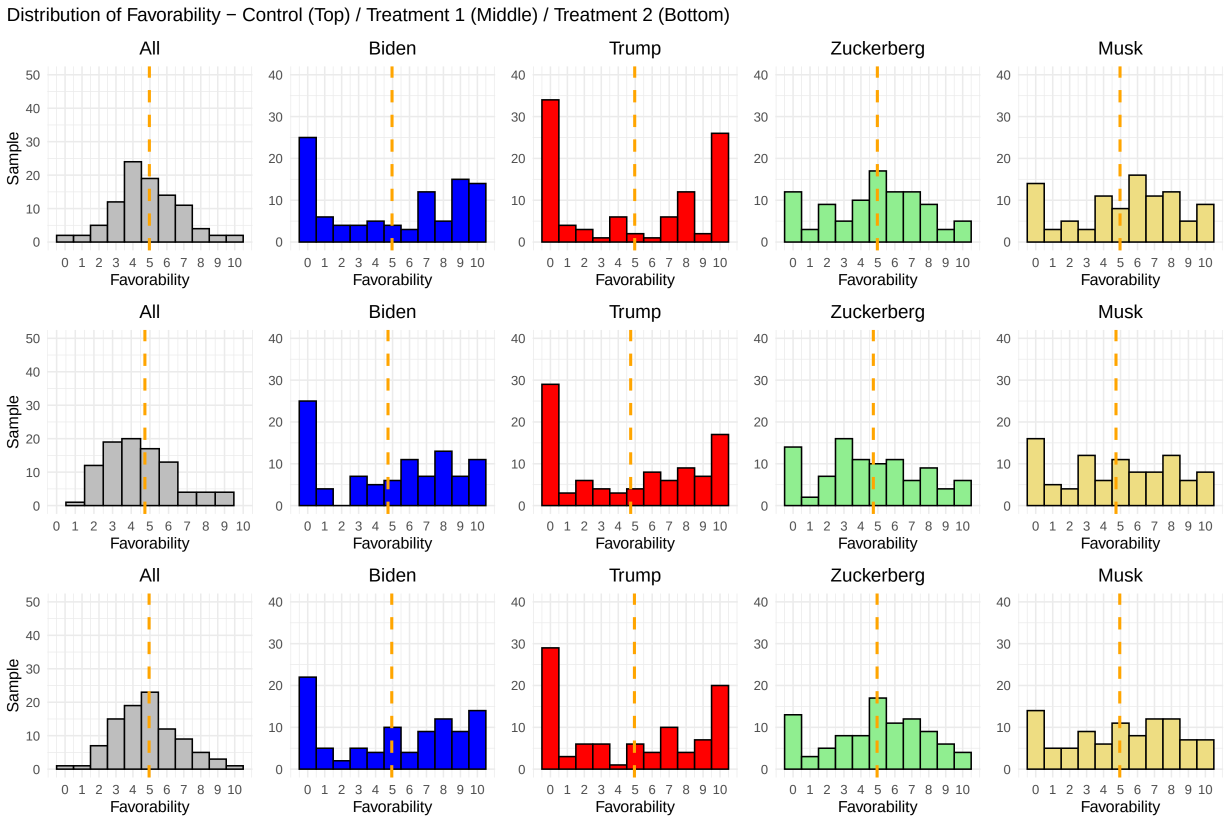

Figure 5: Distribution of Favorability for All Treatment Groups (Control: Top, Treatment 1: Middle, Treatment 2: Bottom)

average favorability scores as well as the individual favorability scores, we are able to estimate the average treatment effect for the average and across the individuals. The results are presented in Table 1.

Table 1: Baseline Regression Results: Average and Individual Favorabilities

	Dependent variable:				
	Favorability All	Favorability Biden	Favorability Trump	Favorability Zuckerberg	Favorability Musk
	(1)	(2)	(3)	(4)	(5)
treatment1	−0.237	−0.115	−0.106	−0.283	−0.445
	(0.282)	(0.544)	(0.594)	(0.416)	(0.459)
treatment2	−0.016	0.166	−0.085	0.092	−0.237
	(0.277)	(0.549)	(0.598)	(0.412)	(0.454)
Constant	4.969***	4.979***	4.835***	4.804***	5.258***
	(0.200)	(0.397)	(0.435)	(0.287)	(0.316)
Observations	289	289	289	289	289
R^2	0.003	0.001	0.0001	0.003	0.003
Adjusted R^2	−0.004	−0.006	−0.007	−0.004	−0.004
Note:	*p<0.05; **p<0.01; ***p<0.001 Robust standard errors are in parentheses.				

For most coefficients, the observed relationships align with our expectation that both treatment groups would show a decrease in favorability. However, the results indicate that neither variable is statistically significant enough to determine, with 95% confidence, that they indeed reduce favorability. Specifically, the first column shows the average favorability across the four individuals tested. While treatment without the disclaimer has a coefficient of -0.237, this result has a p-value of greater than 0.05. For treatment with the disclaimer, the coefficient is -0.016 also with a p-value greater than 0.05. Therefore, when considering the average and individual favorabilities of the four individuals, we fail to reject our hypothesis that AI content has a negative effect on favorability. Further, as neither treatment groups have significance, the isolated effect of the disclaimer is also not significant, which is validated by running the model on treatment 1 vs treatment 2 (not shown in table).

Importantly in Table 1, we see an R^2 value of 0.003 in the average favorability and similar value across the individuals, suggesting that the residual variance may be high. In an attempt to explain the residual variance and address potential omitted variable bias, we build onto the linear model to test the relationship with covariates in the data. We build upon the baseline model with average favorability as the dependent variable by adding each demographic variable (gender, age group, political party, political ideology, and industry) one at a time. With a series of F-tests, we compare the model fit to determine if adding a covariate improves the model. Additionally, we find that there are moderately high correlations between the covariates and political party and political ideology, thus we remove political ideology to avoid multicollinearity. The final improved multivariate model turns out to be:

$$averagefavorability = treatment + agegroup + gender + politicalparty + industry$$

Regression results with the improved model are shown in Table 2. Starting with the treatment groups, the results indicate that neither Treatment 1 nor Treatment 2 significantly affect the overall favorability scores, as shown by the lack of statistical significance in their coefficients across all models. The treatment effects on favorability are small and slightly varied across the different figures. However, these effects are not statistically significant, suggesting that the AI-generated content does not substantially influence participants' perceptions.

Table 2: Multivariate Regression Results: Average and Individual Favorabilities

	Dependent variable:				
	Favorability All	Favorability Biden	Favorability Trump	Favorability Zuckerberg	Favorability Musk
	(1)	(2)	(3)	(4)	(5)
treatment_groupTreatment1	−0.371	−0.488	0.030	−0.530	−0.498
	(0.254)	(0.413)	(0.471)	(0.394)	(0.419)
treatment_groupTreatment2	−0.087	0.154	−0.191	−0.021	−0.290
	(0.254)	(0.428)	(0.489)	(0.401)	(0.423)
age_group35-49	−0.937**	−1.081*	−0.161	−1.671***	−0.833
	(0.291)	(0.488)	(0.520)	(0.446)	(0.462)
age_group50-64	−0.736*	−0.695	−0.824	−0.950*	−0.474
	(0.312)	(0.458)	(0.535)	(0.444)	(0.480)
age_group65+	−1.579***	−0.887	−1.845***	−2.089***	−1.494**
	(0.276)	(0.489)	(0.543)	(0.429)	(0.483)
genderMale	0.462*	0.452	0.116	0.540	0.741
	(0.235)	(0.389)	(0.433)	(0.359)	(0.383)
genderNon-binary / third gender	−0.809**	−1.299**	−4.198***	−0.169	2.428***
	(0.298)	(0.458)	(0.546)	(0.428)	(0.477)
partyRepublican	0.496*	−5.526***	5.413***	−0.438	2.533***
	(0.238)	(0.368)	(0.431)	(0.387)	(0.408)
partyIndependent	−0.883**	−3.654***	1.100*	−1.549***	0.572
	(0.284)	(0.446)	(0.545)	(0.400)	(0.471)
partyDecline to answer	−1.360	−3.010*	−0.068	−1.785	−0.578
	(1.046)	(1.360)	(1.699)	(1.057)	(1.148)
industry_techTechnology	0.725	1.076	0.511	0.937	0.376
	(0.429)	(0.655)	(0.692)	(0.589)	(0.575)
Constant	5.563***	8.440***	3.068***	6.248***	4.497***
	(0.293)	(0.450)	(0.582)	(0.421)	(0.469)
Observations	289	289	289	289	289
R^2	0.204	0.439	0.396	0.147	0.178
Adjusted R^2	0.173	0.417	0.372	0.113	0.146
Note:	*p<0.05; **p<0.01; ***p<0.001 Robust standard errors are in parentheses.				

Based on the results, we also gain an important understanding of the covariates' effect on favorability. Age is a significant predictor of favorability scores, particularly for the oldest age group (65+), which consistently shows a significantly lower favorability compared to the younger baseline age group across the board. This suggests that older participants, especially those aged 65 and above, may be more negatively impacted by the AI-generated content, possibly due to less familiarity with online platforms and modern technology. Political party affiliation yields results in line with the known political leanings of public figures. While being a Republican, as opposed to a Democrat, is associated with a higher favorability towards Trump and Musk, it shows a lower Biden's favorability. Likewise, Democrats show a significantly higher favorability towards Biden. Gender differences are evident, with males generally showing higher favorability, though not always statistically significant. While the "Non-binary / third gender" category appears to show a strong negative effect on Trump favorability, this category is represented by only a single participant, limiting the reliability of these results and should be interpreted with caution due to the small sample size. The industry category included is divided into two groups: the technology industry and non-tech industries. The regression results show that being part of the tech industry has a generally positive effect on favorability, however, these effects are not statistically significant. Overall, the results suggest that demographic factors such as age and political affiliation have a more pronounced impact on favorability ratings than the AI-generated content itself.

We isolate the effects of the treatment variables for the key demographic variables by running the model separately, filtering on each subgroup of demographics. The findings consistently reveal that while treatment does not significantly affect overall favorability, the age group 65+ exhibits a significant difference between Treatment 1 and Control group, as shown in Table 3. The highly significant result is supported by our reasoning that seniors are more susceptible to misinformation due to lower familiarity with digital media and technology, making them more likely to be affected by AI-generated content. Yet, there is no significant effect for older people in treatment 2, signifying that adding a disclaimer that the content is fake can help alleviate the negative effect of the videos on their perception of the public figures. Our results show that although it may seem that AI content has no effect on shaping public opinion, AI content has more nuanced effects for people of different ages and that disclaimers still play an important role in minimizing the effect of deepfake content.

Table 3: Regression Results by Age Group

	Dependent variable:			
	Average Favorability			
	18-34	35-49	50-64	65+
	(1)	(2)	(3)	(4)
Treatment 1	−0.390	0.217	−0.929	-1.033^{**}
	(0.486)	(0.469)	(0.639)	(0.364)
Treatment 2	0.192	0.136	−0.390	−0.581
	(0.579)	(0.557)	(0.516)	(0.462)
genderMale	0.726	0.846	0.535	−0.032
	(0.507)	(0.503)	(0.476)	(0.422)
genderNon-binary / third gender	−0.521			
	(0.446)			
partyRepublican	0.635	0.804	0.643	0.236
	(0.465)	(0.489)	(0.548)	(0.385)
partyIndependent	−0.912	−0.704	−1.116	−0.275
	(0.517)	(0.607)	(0.646)	(0.401)
partyDecline to answer	-3.354^{***}	0.743		
	(0.547)	(0.511)		
industry_techTechnology	1.192	1.096	-1.522^{*}	0.688
	(0.656)	(0.772)	(0.761)	(1.359)
Constant	5.323^{***}	3.916^{***}	5.257^{***}	4.532^{***}
	(0.509)	(0.418)	(0.564)	(0.384)
Observations	88	69	71	61
R^2	0.299	0.213	0.148	0.134
Adjusted R^2	0.228	0.123	0.069	0.038

Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$
Robust standard errors are in parentheses.

Heterogeneous Treatment Effects

Using a fully saturated model can lead to a better-specified model that allows us to study the nuanced relationships in the data by adding the interaction terms between the covariates. This reduces the residual variance, resulting in lower p-values for predictors across the board. In our fully saturated model, average favorability is regressed on the treatment indicators, demographic covariates, and their interaction terms. The significant coefficients are shown in Table 4.

Notably, the p-value of the 65+ age group increases to the point where it is no longer significant in the model. This may be attributed to omitted variable bias in the simpler model, which may overestimate the effect of older age. In the more complex model, some of the treatment effect for this age group is explained by another variable or interaction, reducing the coefficient.

Several other coefficients make sense to us intuitively. For example, the coefficient of the Technology value in the job industry variable shows us that people who work in the tech industry rate the public figures around 2.6 points higher on average regardless of treatment. This makes sense since there are two tech CEOs, and presumably a person working in tech is more likely to rate the two entrepreneurs higher in favorability.

The significant interaction terms suggest a change in average favorability when a person is within two demographic groups. Take a male in the tech industry for example. This person rates all the figures lower on average by almost 2 points. It is difficult to interpret this result, especially since it contradicts the narrative of the coefficient of the Technology value, but one possible explanation is that specifically, males working for tech companies that compete with Tesla or Meta may rate the CEOs lower regardless of whether they are in control or treatment. However, this is a conjecture that needs to be experimented on to make any further claims. Our fully saturated model presents intriguing findings, but it is important to note that there are no heterogeneous treatment effects. We only find some significant covariates. None of them are conclusive or causal, but they do suggest a direction for future research.

Table 4: Significant Coefficients from the Interaction Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.035	0.593	8.486	0
age_group35-49	-1.707	0.746	-2.289	0.023
genderMale	1.472	0.691	2.129	0.034
industry_techTechnology	2.595	1.158	2.241	0.026
age_group35-49:partyDecline to answer	5.488	2.304	2.382	0.018
genderMale:industry_techTechnology	-1.897	0.828	-2.291	0.023

CACE Analysis

In addition to regression analysis, we calculate noncompliance to account for never-takers in our experiment. Never-takers are participants who skipped or ignored the treatment videos. While in other scenarios always-takers also belong in the noncomplier category, we dismiss the possibility of participants seeing our created videos outside of the experiment or participants discussing the survey amongst themselves. To address never-takers, we calculate the Complier Average Causal Effect (CACE), using the Intent to Treat (ITT) effect as our baseline. The ITT effect can be biased if never-takers skew the results as they might dilute the treatment effect.

We define compliers as those who took at least 102 seconds to complete the survey, based on the fact that the treatment videos totaled 62 seconds and the survey requires an estimate of additional 40 seconds to complete. Using this criteria, we find a 94.3% compliance rate for Treatment 1. Dividing the ITT effect by this take-up rate, we calculate a CACE of -0.2517, which is also not a significant treatment effect. This analysis affirms that the treatment does not significantly affect favorability scores, consistent with our regression and t-test findings.

Limitations and Next Steps

In our study, several limitations warrant discussion. First, while our sample size of 289 participants provides a solid foundation, a larger sample would enhance the statistical power, enabling us to detect smaller effects and increase the generalizability of our findings. Additionally, we attempt to utilize standardized treatments across all videos, but it is difficult to create uniformly negative yet distinct scripts for each public figure. Noncompliance is another concern, as some participants may have yet to fully adhere to their assigned treatment conditions by failing to watch the provided videos. Despite our efforts to account for this, noncompliance could introduce bias. Future research should address these limitations by increasing sample size, exploring varied content, and implementing more robust compliance and checks.

Implications and Conclusion

In running the experiment to test the effect of showing AI-generated deepfake videos on participants' favorability of famous public figures, our results indicate that overall, the videos do not have a significant effect. While deepfake videos seem to directionally decrease favorability, they do so within a margin of error. Our null hypothesis that treatment will have no effect on favorability fails to reject. The effect of the disclaimer in treatment 2 is also not significant. However, we find a significant result in the 65+ age group. This indicates that older people may be more susceptible to the harmful effects of artificial videos.

We acknowledge that our experiment may not be capturing the full effects of AI. Longer and more constant exposure to fake and misleading content may have effects that are unmeasured in this experiment. The videos we create for this experiment are short with limited functionality. As artificial intelligence grows in capability, its effectiveness will only increase as technology improves. As this happens, the use of AI and deepfake content must be done with caution, especially as older people may be affected more easily than younger people.

Citations

Bontridder, N., Poullet, Y. (2021). The role of artificial intelligence in disinformation. https://www.cambridge.org/core/journals/data-and-policy/article/role-of-artificial-intelligence-in-disinformation/7C4BF6CA35184F149143DE968FC4C3B6

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social Media + Society. https://doi.org/10.1177/2056305120903408

Wittenburg, C., Epstein, Z., Berinsky, A.J., Rand, D.G. (2023). Labeling AI-Generated Content: Promises, Perils, and Future Directions. MIT Schwarzman College of Computing. https://computing.mit.edu/wp-content/uploads/2023/11/AI-Policy_Labeling.pdf#:~:text=URL%3A%20https%3A%2F%2Fcomputing.mit.edu%2Fwp