AI Labeling

# Study 1

We examine whether people are influenced by the content of AI-generated (“deepfake”) videos and whether labeling the video as AI-generated reduces this influence. Participants are asked to rate their opinions on whether AI should be regulated by the government using a 6-item scale. They then watch a video presenting an argument either in favor (“Pro”) or against (“Anti”) such regulation. After watching, participants report their opinions again using the same scale. Half of the participants are informed that the video is AI-generated before viewing, and an “AI-generated” label appears on the corner of the video. The other half are not given this information. We assess the impact of the video by comparing participants’ pre- and post-video opinions. We predict that watching a Pro (Anti) deepfake video will increase (decrease) participants’ support for AI regulation, and that the AI label will not eliminate this effect.

## Methods

We used ChatGPT to generate two transcripts for a 4-minute speech: one advocating for government regulation of AI and the other opposing it. These transcripts were pretested for their effectiveness in influencing opinions on the issue. A professor from a northeastern U.S. university delivered both speeches on camera. To ensure consistency, all videos were recorded in the same setting, with the professor maintaining a consistent appearance. Using the real Pro-regulation video, along with additional footage and audio of the professor, we trained an algorithm to generate a deepfake version of the Anti-regulation speech. Likewise, we created a deepfake Pro-regulation video using the real Anti-regulation speech as the base. This process resulted in four videos: two real recordings (Pro and Anti) and two deepfake versions. Additionally, we produced labeled versions of the deepfake videos by overlaying “AI-generated” in the upper left corner. In total, we created six videos.

We recruited 1,802 participants ( = 42.52 years; 47.8% Male) from Prolific. They first completed a six-item scale assessing their opinions on government regulation of AI. Sample items included: “Government regulation is necessary to ensure that generative artificial intelligence is used ethically,” “The government should impose strict regulations on the development and deployment of generative artificial intelligence,” and “Government regulation will stifle innovation in generative artificial intelligence.” Participants used sliders ranging from -100 (“Strongly disagree”) to 100 (“Strongly agree”) to report their opinions. Next, participants were randomly assigned to watch one of the six videos. They were encouraged to watch carefully, as they could earn a bonus by answering three content-related questions, receiving $0.10 for each correct response. Immediately after watching the video, participants completed the same six-item scale on AI regulation, with slider defaults set to their initial responses. Additionally, participants were asked to rate how persuasive they found the video on a 5-point scale. They then answered three bonus questions related to the video. Finally, participants indicated whether they thought the video was AI-generated (yes or no). The survey concluded with basic demographic questions.

## Results

We begin by examining how the videos influenced participants’ opinions. Before and after watching the video, participants reported thier opinions on AI regulation on a 6-item scale twice (with 3 items reverse coded; =0.86). Their average support for regulation is = 25.1 and = 22.83, respectively, on a scale ranging from -100 to 100.. We define “Change in Direction” as the difference between post-watch and pre-watch opinions (Post-watch opinion - Pre-watch opinion). This variable is reverse-coded to account for the video’s stance, ensuring that an increase in support after watching a Pro-regulation video and a decrease in support after watching an Anti-regulation video both reflect positive changes. We find that participants who watched a deepfake video, knowing it was AI-generated, adjusted their opinions by an average of 10.02 points. This suggests that labeling an AI-generated video does not prevent it from influencing participants’ opinions, , ).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Column 1 shows the Change of Opinion in Direction based on whether the video is deepfake, whether a label is presented, and whether the video is Pro regulation. Column 2 displays the perceived persuasiveness based on whether the video is deepfake, whether a label is presented, and whether the video is Pro regulation   |  |  |  | | --- | --- | --- | | Deepfake | -2.854+ | -0.313\*\*\* | |  | (1.457) | (0.069) | | Label Shown | -1.392 | -0.099 | |  | (1.451) | (0.068) | | Pro Regulation | -6.374\*\*\* | 0.319\*\*\* | |  | (1.187) | (0.056) | | Constant | 17.734\*\*\* | 3.040\*\*\* | |  | (1.208) | (0.057) | | N | 1802 | 1802 | | * p < 0.1, \* p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001 | | | |

Next, we examine whether adding a label reduces the impact of deepfake videos. Column 1 of Table [Table 1](#tbl-study1regs) presents the results of a linear regression model with “Change in Direction” as the outcome variable. The model includes predictors for whether the video is a deepfake, whether it is labeled, and whether the video is Pro-regulation (to assess whether the impact differs across conditions). We find that speeches in real recorded videos significantly shift participants’ opinions, while deepfake videos with the same content have a marginally weaker effect.Importantly, as predicted, labeling a video as AI-generated does not reduce its influence on participants’ opinions. This finding is consistent when participants directly rate the video’s persuasiveness—adding a label does not reduce how persuasive they find the content. Interestingly, while videos arguing against regulation influence opinions more, participants mistakenly perceive Pro-regulation videos as more persuasive. This may be because the content of Pro-regulation videos aligns more closely with their prior beliefs.

These findings are not due to participants failing to notice the label, as they responded to it when identifying whether the video was AI-generated. Among those who watched a deepfake, 78% believed the video was AI-generated when no label was present, compared to 90% when the label was shown. In other words, adding a label significantly increased participants’ awareness of the video’s origin, (1, n = 1,200) = 34.2, p < .001, but had no impact on the video’s influence.

## Discussion

Contrary to the expectations of AI policy, labeling a video as AI-generated does not reduce its influence on viewers. Even when people are aware that the person in the video did not actually say the words presented, their opinions are still swayed in the direction of the video’s message.

# Study2

In Study 1, we randomly assigned participants to watch different videos and found that AI labeling does not prevent people from being influenced by AI-generated content. In the next study, we allow participants to choose which video to watch. We predict that labeling a deepfake video as AI-generated will spark curiosity, making participants more likely to select it. As a result, increased exposure may backfire, drawing more attention to the content and amplifying its influence.

## Methods

We used the real recorded video arguing for AI regulation and two versions of deepfake video aruging against regulation (with and without label) created in Study 1. We recuited 999 participants ( = 47.45 years; 48.1% Male) from Prolific and asked whether they would like to watch a video in which a professor argues for AI regulation or one in which the professor argues against it. Half of the participants were assigned to the “No Label” condition, where they made their choice based solely on the video’s content, without knowing how it was created. The other half were in the “Label” condition, where they were informed that the video opposing regulation was AI-generated before making their choice. Before watching their selected video, participants were told they could earn $0.10 for each correct answer on three bonus questions about the video’s content, encouraging them to watch carefully. After viewing the video, they reported their opinions on AI regulation using the same six-item scale from Study 1, now measured on a 7-point scale ranging from strongly disagree to strongly agree. Finally, they answered the three bonus questions, and the survey concluded with basic demographic questions.

## Results

We first examine whether labeling a deepfake as AI-generated increases its appeal. When participants were unaware that the video was AI-generated, 34% chose to watch the video opposing regulation. However, when informed that the video was created using artificial intelligence, this number increased to 43%. This suggests that AI labeling makes the video more appealing, increasing its likelihood of being selected, (1, n = 999) = 6.96, p = 0.008.

We then compare participants’ opinions on whether AI should be regulated across conditions. On average, participants in the “No Label” condition reported a support level of 4.81 out of 7 aafter watching their selected video, while those in the “Label” condition reported a support level of 4.92. Contrary to our prediction, this difference was not statistically significant (, ). One possible explanation is that some participants who strongly supported AI regulation chose to watch the deepfake anti-regulation video out of curiosity but were resistant to changing their stance.

When examining only participants who watched the anti-regulation videos, we found that knowing the video was AI-generated reduced its persuasive impact, , . Among this subset, 82% of participants in the “No Label” condition believed the video was AI-generated, compared to 92% in the “Label” condition, who correctly identified it as AI-generated. This might suggest that when participants mistakenly believed the video was real, they were more influenced by its content.