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 ($M_{\text{Age}} = 42.52$ 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; α =0.86). Their average support for regulation is $M_{\text{beliefPre}}$ = 25.1 and $M_{\text{beliefPost}}$ = 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, $t(611) = 10.43$, $p < .001$).

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

	(1)	(2)
Deepfake	-2.854+	-0.313***
	(1.457)	(0.069)
Label Shown	-1.392	-0.099
	(1.451)	(0.068)

	(1)	(2)
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 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, $\chi^2(1, n = 1{,}200) = 34.2$, $p < .001$, 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 2, we shift our focus from influence to exposure, which is a necessary precursor to any persuasive effect. Rather than assigning content, we allow participants to choose which video to watch. This design enables us to examine whether AI labeling influences participants' willingness to engage with the content. Specifically, we test whether labeling a deepfake as AI-generated increases or decreases viewers' interest in watching it, thereby affecting the likelihood of exposure.

Methods

We recruited 1,545 participants (M_{Age} = 45.01 years; 47.7% Male) from Prolific. Participants first reported their beliefs on AI regulation using the same 6-item scale as in Study 1. We categorized them into two groups based on their responses: those who supported regulation (scores above 0) and those who opposed it (scores below 0). We randomly assigned those with score 0 into one of these groups. Participants then choose between watching a video arguing for AI regulation or one arguing against it. Indeed, the video that opposed to participants' beliefs was AI-geenrated, and the one aligns with their beliefs was real-recorded. This setup allows us to examine the effect of AI labeling on participants' willingness to watch a video controlling for the desire to watch content that aligns with their beliefs. Half of the participants were informed that the video that goes against their beliefs was AI-generated, creating using deepfake technology, putting the words in the professor's mouth. Participants then watched their selected videos, and reported their opinions on AI regulation again using the same six-item scale. Finally, they answered three bonus questions related to the video's content and provided demographic information.

Results

To our most interest is how likely particiapnts were to choose the deepfake video that argued against thier beliefs. Labeling the video as AI-generated significantly decreased the likelihood of participants choosing to watch it, $\chi^2(1, \text{n} = 1{,}545) = 11.09$, $\text{p} < .001$. Specifically, 39% of participants chose to watch the belief-opposing video when it was not labeled as AI-generated, compared to 31% when it was labeled. Participants tend to avoid content that is created using deepfake technology and more willing to receive content that is real-recorded, when the content is against their beliefs.

This pattern held regardless of the participants' original stance on AI regulation. Among 75% of participants who supported AI regulation ($M_{\text{beliefPre}}$ = 38.43, SD = 27.1), 30% chose to watch the anti-regulation video when it was not labeled as AI-generated, compared to 25% when it was labeled. Similarly, among the 25% who opposed AI regulation ($M_{\text{beliefPre}}$ = -21.28,

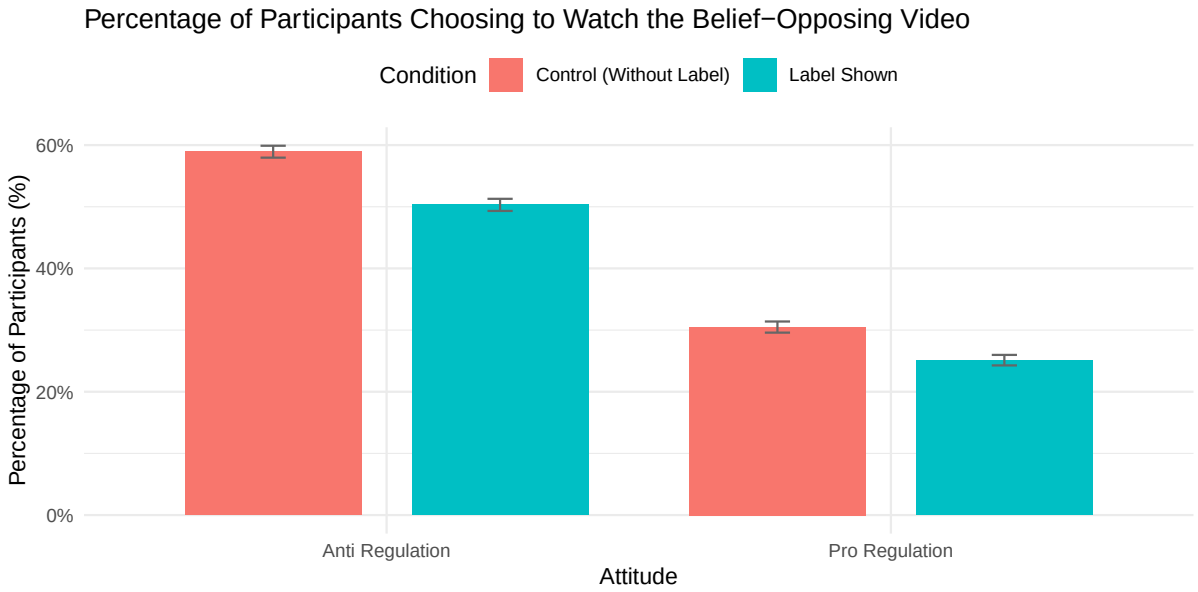


Figure 1: Video choice in Study 2. The bar chart shows the percentage of participants who chose to watch the video that argues against their beliefs, by condition and attitude.

SD = 23.36), 59% selected the pro-regulation video when it was unlabeled, versus 50% when it was labeled (see Figure 1).

Next, we compare the change in beliefs after watching the video across conditions. Specifically, we examine the extent to which participants' opinions on AI regulation shift in the opposite direction of their initial stance (that is, becoming less supportive among those who initially supported regulation, and less opposed among those who initially opposed it). We compare the belief reported after watching with their initial beliefs, and define changes that move toward the opposite direction as positive changes. We find that the change in opinions is significantly different between the two conditions, with participants seeing the AI labels showing less change compared to those who were not informed how the video opposing their belief was created, $t(1543) = -3.04$, $p = .002$. This difference may be driven by the fact that participants avoid content that is created using AI technology, and thus are less likely to watch the deepfake video that contradicts their beliefs, and therefore less likely to change their opinions.

We then focus on the subset of participants who watched deepfake videos, with or without labels, and examine whether they were influenced by the content. Regardless of their initial stances, among those who chose to watch the video created by AI (that is, the video that goes against their opinions), including an AI label does not prevent participants from being influenced, as there is no significant difference in their change in opinions between conditions, $t(533) = -1.65$, $p = .100$. This replicates our findings in Study 1, suggesting that labeling a video as AI-generated does not eliminate its persuasive effect.

Discussion

When participants are given a choice, they are less likely to watch a video that argues against their beliefs if it is labeled as AI-generated versus not knowing how it is created. This findings suggest that labeling a deepfake as such reduces people's willingness to voluntarily exposure, making its content less appealing. However, as shown in Study 1, AI label does not prevent people from being influenced by the content when they are exposed to it. AI label may help reducing the influence of content created by AI in reducing its exposure, but not its influence once people choose to watch it[1].

Study3

So far, we know that at the personal level, AI labeling does not reduce the persuasive effect of deepfake content but does reduce the willingness for voluntary exposure. Finally, we examine the impact of AI labeling on the willingness to share deepfake videos, focusing on how it influences the dissemination of deepfake content.

Methods

We recruited 142 participants ($M_{\text{Age}} = 41.54$ years; 45.1% Male) from Prolific. Participants first reported their beliefs on AI regulation using the same 6-item scale as in Study 1, based on which we categorized them as supporting regulation (score above 0) or opposing regulation (score below 0). They then watched two one-minute video clips arguing for and against AI regulation, respectively. They were told that these clips were excerpts from a four-minute speech. The video opposing their beliefs was a real recording of a professor, while the video aligning with their beliefs was a deepfake version of the same professor. Participants were randomly assigned to one of three conditions: "Label," "Informed," or "No Label." In the "Label" condition, participants were told that the video opposing their beliefs was AI-generated, and a label reading "AI generated" was displayed in the upper left corner throughout the video. In the "Informed" condition, participants were told the video was AI-generated but no label was presented in the video. In the "No Label" condition, they were not given this information. We randomized the order in which the real-recorded and deepfake videos were played to control for order effects.

After watching both videos, participants chose which video they preferred to share with another person. We told them that we would show a full four-minute speech to another Prolific user, and they could decide which video to share. Importantly, in the "Label" and "Informed" conditions, participants themselves knew which video was AI-generated, but in the "Informed"

[1]We conducted a similar study in 2024, where participants were also given the chance of choosing which video to watch. It was a year before we ran Study 2 and at the time deepfake videos were not as common as they are now. Please refer to Appendix for the details of the study.

condition, the AI label was not shown, so the other user would not know the video's origin. After making their choice, participants reported their opinions on the two video clips. They rated the extent to which sharing each video would influence others' opinions, mislead others, cause confusion, and represent a responsible way to share information, using a five-point scale from "Not at all" to "Extremely." We also asked how they believed others would perceive them if they shared the video, specifically how knowledgeable, humorous, trustworthy, lacking in judgment, and trendy they would appear, on a five-point Likert scale ranging from "Not at all" to "A great deal." We then explored participants' reasons for choosing a particular video to share. We reminded them of their choice and asked to what extent each of the following factors influenced their decision: the video presented strong arguments, was enjoyable, was factually accurate, presented arguments that are less commonly heard, had high video quality, or was chosen at random. Finally, participants provided demographic information.

Results

We first look at whether participants were willing to share the deepfake video that argued against their beliefs. When participants were not informed that the video was AI-generated, 73% chose to share it. However, when they were informed that the video was AI-generated, this number dropped to 42% in the "Informed" condition and 31% in the "Label" condition. This suggests that labeling a deepfake as AI-generated reduces its appeal, making it less likely to be shared, $\chi^2(2, \text{n} = 142) = 18.53$, $\text{p} < .001$.

Next, we examine participants' perceptions of the two videos. Please see Figure 2 for how participants perceived sharing the videos across conditions. When participants did not know how the video was created, they believed sharing the real video that aligned with their own stances would mislead others' opinions less, cause less confusion, and be a more responsible way to share information. In contrast, sharing the deepfake video was perceived as more misleading, more confusing, and less responsible. Interestingly, when participants were informed that the video aligning with their beliefs was AI-generated, their perceptions shifted. They saw sharing it as more misleading and confusing and less responsible. This suggests that AI labeling makes participants more cautious when considering sharing AI-generated content.

In addition to their personal opinions on sharing, how participants anticipated others would view them when sharing deepfake content might also influence their decisions. When not told how the video was created, participants felt that sharing a video aligning with their beliefs would make them seem more knowledgeable, $t(48) = -1.76$, $p = .085$, though also somewhat lacking in judgment, $t(48) = 1.85$, $p = .070$, with no significant differences in perceived trustworthiness, trendiness, or humor. However, when they were explicitly informed that the video aligning with their beliefs was created using deepfake technology, they believed that sharing it would make them appear more negative. In the "Label" condition, where others would also see that the shared video was created using AI, participants believed that sharing the deepfake video would make them appear less trustworthy, $t(44) = 3.15$, $p = .003$, more lacking in judgment, $t(44) = -2.30$, $p = .026$, and less knowledgeable, $t(44) = 2.67$,

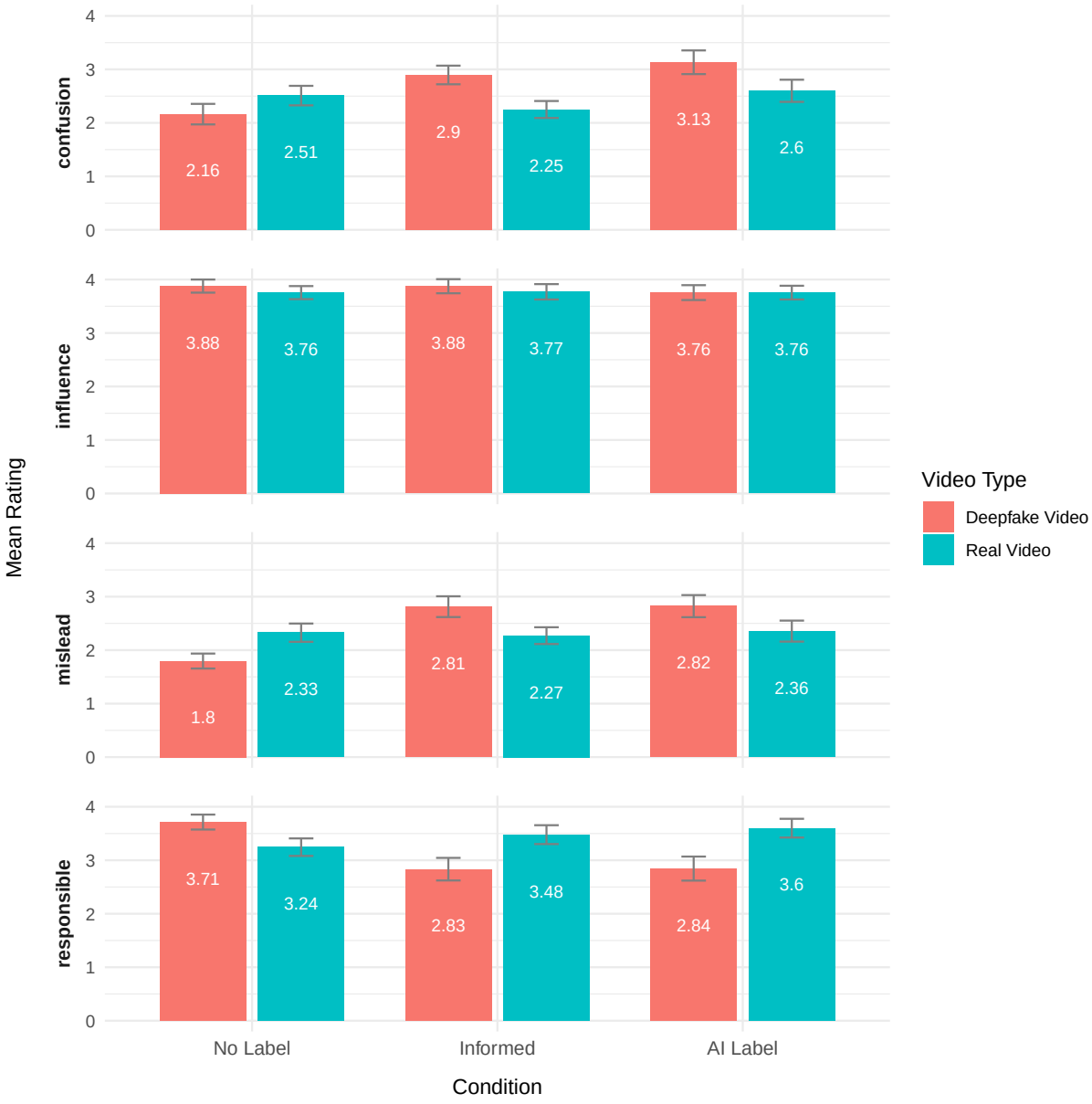


Figure 2: Participants' perceptions of sharing the videos in Study 3. The bar chart shows how participants viewed sharing the videos, separated by condition and video type.

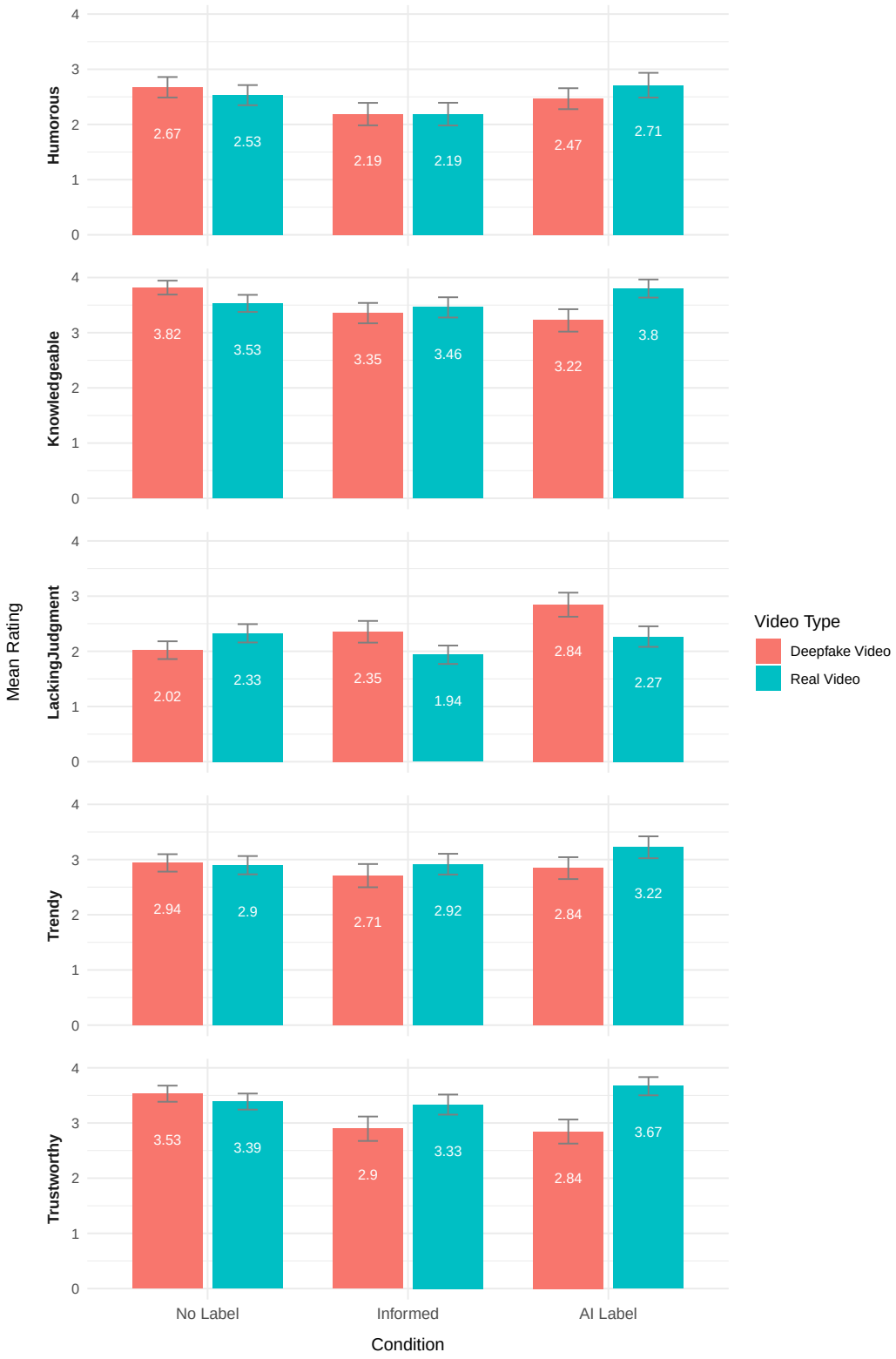
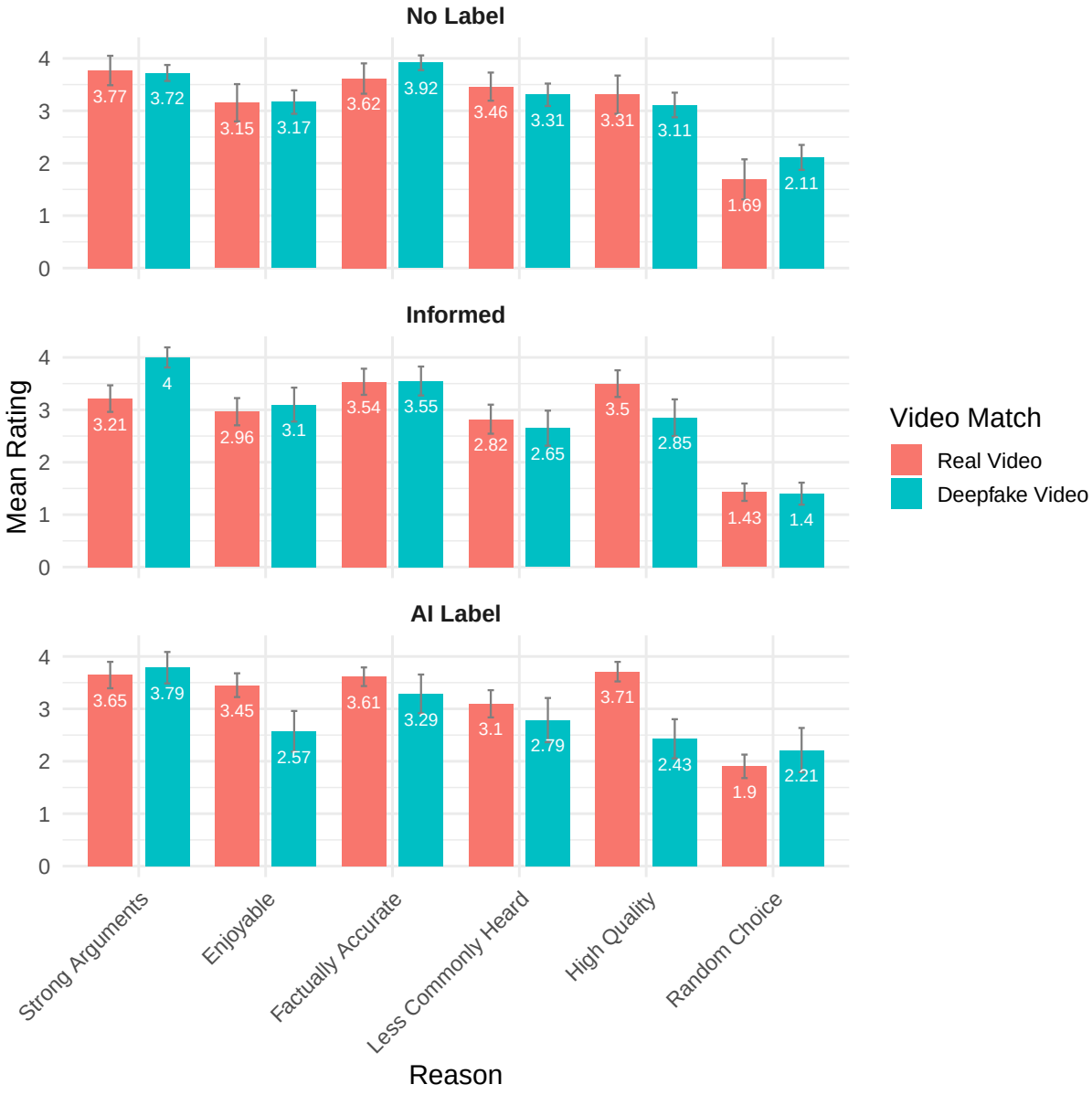


Figure 3: Anticipated impressions from sharing each video in Study 3. The bar chart shows how participants believe others would perceive them, broken down by condition and video type.

$p = .011$, compared to sharing the real video. This suggests that AI labeling might reduce the willingness to share deepfake content by harming the sharer's social image. A similar but smaller pattern emerged in the "Informed" condition, where participants knew the video was AI-generated but no label was shown: sharing the deepfake was still viewed as making them appear less trustworthy, $t(47) = 1.93$, $p = .060$, and more lacking in judgment, $t(47) = -1.87$, $p = .067$, though the differences from the real video were less pronounced. Figure 3 fpresents the anticipated impressions of sharing real and deepfake videos across conditions.



Lastly, we examine participants' reasons for choosing to share a particular video. Across all

conditions, participants rated "the video presented strong arguments" as the most influential factor in their decision, followed by "the video was factually accurate" and "the video was enjoyable." Notably, participants who were not informed that the video aligned with their beliefs was created by AI reported sharing it due to its accuracy. In the "Label" condition, participants decided to share the real recorded video for the very same reason even though it went against their stance. Please see **?@tbl-study3reasons** for the mean ratings of various factors influencing participants' decisions to share a video, separated by condition and video type.

Appendix

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 (M_{Age} = 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, $\chi^2(1, \text{n} = 999) = 6.96$, $\text{p} = 0.008$.

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 ($t(997) = 1.27$, $p = .206$). 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, $t(384) = 3.40$, $p < .001$. 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.