Labeling Deepfake Videos Reduces Exposure But Not Persuasiveness

Amanda Chen[1], David Hagmann[1], Christie Pang[2], and George Loewenstein[3]

[1]Department of Management, The Hong Kong University of Science and Technology

[2]Division of Emerging Interdisciplinary Areas, The Hong Kong University of Science and Technology

[3]Department of Social and Decision Sciences, Carnegie Mellon University

Author Note

The authors declare no conflicts of interest. GL initiated the research question and all authors contributed to the design of the studies. CP generated the deepfake videos. AC programmed the survey experiments and conducted the analyses. DH replicated the analyses. AC and DH wrote the first draft of the manuscript and all authors edited and approved the final version.

We are most grateful to a colleague for recording the arguments in favor and against regulation and allowing us to generate deepfake videos based on those recordings. The arguments were generated by ChatGPT and do not necessarily reflect his views. This research is partially funded by the Hong Kong Research Grants Council Early Career Scheme under Award Number 26506323, the HKUST Institute for Emerging Market Studies with support from EY, and internal funding from Carnegie Mellon University.

Correspondence concerning this article should be addressed to Amanda Chen, Email: zchengj@connect.ust.hk

Abstract

Increasingly realistic AI generated deepfake videos are raising alarm about their potential use in disinformation campaigns. Viewers may be misled into believing the depicted content is real and thus get persuaded on policy-relevant beliefs. A commonly proposed solution is to require disclosure, such as via text embedded in the video that informs the audience that the content is AI generated. Across four large scale experiments (N = 7,107), we show that the concern about persuasion is justified, but that disclosure is only partially effective. Specifically, participants who watched realistic but AI generated videos in which a speaker provided arguments related to regulation of AI technology were persuaded regardless of whether it included a warning label, which did not diminish the video's persuasiveness. Viewers rated labeled deepfakes as less deceptive, but otherwise evaluated them and their senders no differently from unlabeled deepfakes. However, disclosure and labeling are effective at the extensive margin: people were less likely to watch a video that is labeled as a deepfake, and less likely to share it with others. Our findings suggest that labels are effective when users opt-in to viewing the content, or can share it with others—but not in contexts such as news media feeds that don't offer users the opportunity to selectively view, or not view, content.

Keywords: Deepfakes, Misinformation, Disclosure, Persuasion

Labeling Deepfake Videos Reduces Exposure But Not Persuasiveness

Videos and images generated by artificial intelligence have become increasingly realistic, often making it difficult to distinguish them from authentic recordings. Synthetic media that are difficult to distinguish from reality (so-called "deepfakes") pose risks for elections, markets, and national security because they can be deployed at scale to mislead audiences and erode trust in information ecosystems (Chesney & Citron, 2019; Hackenburg et al., 2025; Paris & Donovan, 2019). Recent evidence underscores both their realism and their psychological potency: AI-synthesized faces can be indistinguishable from real faces and even perceived as more trustworthy (Nightingale & Farid, 2022), and synthetic political videos can alter judgments about truth and news credibility (Vaccari & Chadwick, 2020).[1] These concerns have motivated policymakers, platforms, and news organizations to seek policies that can avoid the deception of even highly discerning consumers.

One potential solution relies on disclosure regimes (textual or on-screen labels indicating that content was generated by AI) as a potential remedy. Disclosure is a familiar policy tool. In domains ranging from consumer finance to medicine and journalism, disclosure is frequently mandated on the intuition that informing people will enable them to discount biased or low-quality information (Weil et al., 2013). When recipients are informed about problems related to the source of information, they may become more skeptical, recognize strategic incentives, and subsequently ignore the message and resist attempts at persuasion. Disclosures also promote transparency and accountability norms, which may further reduce naive acceptance of deceptive content. Some jurisdictions and institutions are extending this logic to AI generated content by requiring or encouraging labels that identify deepfakes (for example the European Union's AI Act, which requires watermarking for AI generated or manipulated video, Block, 2024).

However, a large literature in psychology, economics, and law cautions that mandated

[1] Generative AI technology is advancing at a rapid pace. In the referenced work, viewers became skeptical of real content rather than being misled. But the ability to generate plausible artificial video content has rapidly advanced over the past few years and, we suspect, will continue to get harder to distinguish from real content.

disclosure often has limited effects and can sometimes backfire: disclosures may be overlooked due to limited attention or cognitive load, misunderstood, or even license more biased behavior and greater compliance with conflicted advice—for example, when advisees feel obligated to heed disclosed advice or advisers feel morally "absolved" after disclosure (Ben-Shahar & Schneider, 2014; Cain et al., 2005; Loewenstein et al., 2014; Sah et al., 2013). Relatedly, corrections and warnings often fail to fully neutralize misinformation due to continued-influence effects (Lewandowsky et al., 2012). In a related context, informing people that they have been microtargeted has no impact on the persuasiveness of the message (Carrella et al., 2025). Thus, while labels may increase awareness that content is AI generated, it remains unclear whether they meaningfully reduce the persuasive impact of what that content says.[2]

Disclosure of how content was created can reduce persuasion to the extent that it changes how people evaluate the message. That is, they may come to understand that the video itself is not real and then ignore the position that is being presented. However, it is possible that people are nonetheless persuaded because they may still attend to the underlying argument, which is independent of whether the speaker actually held the view inherent in what they are shown as saying. Prior work using written arguments finds that changing attribution of text to a human expert or an AI expert model does not change its persuasiveness (Gallegos et al., 2025), and studies have shown that conversations with large language models can lead to lasting persuasion (Costello et al., 2024; Lin et al., 2025). Indeed, persuasion can occur even when audiences know content is fictional: narratives routinely shift beliefs and intentions, including when they are not presented as originating from a fictitious communicator (Green & Brock, 2000). In one notable example, access to a Brazilian telenovela that featured smaller families reduced fertility, consistent with viewers updating expectations and social norms from fictional portrayals (La Ferrara et al.,

[2] There are multiple ways to label AI generated content; for example distinguishing between process-based labels, which reveal how content is created, and harm-based labels, which highlight its potential to mislead (Wittenberg et al., 2025). We focus on labels that disclose how the content was created to parallel existing and proposed disclosure requirements.

2012). Thus, raising awareness about the AI generated nature of a video may make people view it as fictional, yet the content may remain persuasive. We test whether clearly identified deepfake content remains persuasive in our first study.

Another way labels could affect persuasion is by affecting exposure to the content. If people understand that a video is a deepfake, they may want to avoid watching it (Carney et al., 2024). Thus, even if a deepfake is as persuasive as a real video *conditional on exposure*, people may be less likely to view it. For example, while false claims about COVID-19 vaccines (flagged as misinformation) reduced vaccination intention when viewed, they were viewed less frequently than unflagged content (Allen et al., 2024). However, exposure may have been influenced by Facebook's downranking policy. Labels which simply inform participants that a social media image was generated using AI have only limited impact on participants' willingness to engage with the posts (Wittenberg et al., 2025). On the other hand, people have been found to be more receptive to opposing views when expressed by AI (Lu et al., 2025) and may be curious to see realistic AI generated content due to its novelty and thus may be more likely to watch something labeled as a deepfake (Berlyne, 1960; Loewenstein, 1994). This could then lead to more persuasion overall. We test this experimentally in the context of AI generated video content in Study 2.

Relatedly, people may be less willing to share a deepfake with others, thus also reducing how likely others are to have the chance to view it. Prior work shows that simple prompts that shift attention to accuracy can substantially reduce intentions to share misinformation, and that false content spreads differently from true content in social networks (Pennycook et al., 2021; Vosoughi et al., 2018). Thus, it may be that people second-guess their decision to share a video if they are informed that it is a deepfake. Moreover, people may also hesitate more to share content that they know will be identified as a deepfake to downstream viewers. We test this in Study 3, and recruit a separate sample to evaluate perceptions of those who sent AI generated content (Study 3b).

Our predictions point to two margins on which labels could matter. On the *intensive* margin, labels could reduce the credibility and trustworthiness of the source and thus make the content less persuasive. The prior literature on disclosure, however, suggests that this is unlikely. On the *extensive* margin, labels could reduce the number of people who are exposed to the content via the demand side (willingness to watch the video) and via the supply side (willingness to share the video). This second pathway is relevant in the many cases in which people choose what content to consume and share, such as on social media.

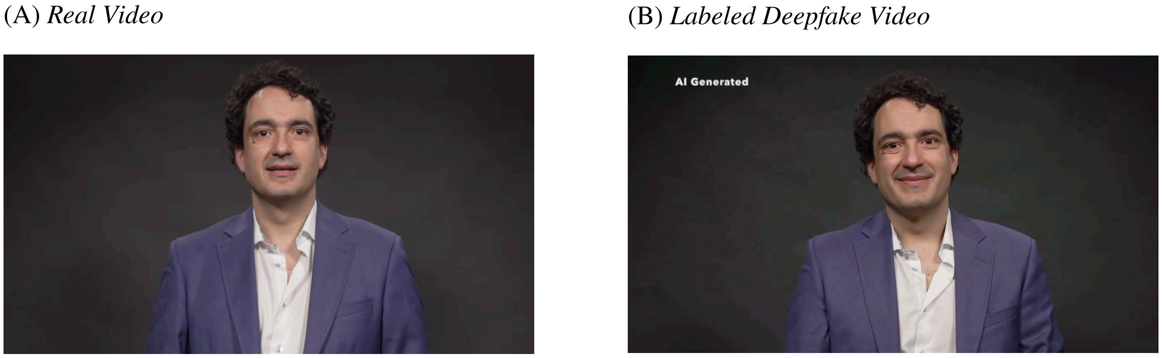
Across four large-scale experiments (N = 7,107), we test whether disclosing how content was created shapes the persuasive impact of AI generated media, the core proposition underlying AI-labeling policies. In Study 1, participants watched a video seeking to persuade them for or against the need for regulating AI technology. They were randomly assigned to watch a real video, an AI generated deepfake, or the identical deepfake but with a label displayed with the video. In Study 2, participants chose whether to watch a video in favor of or against the need for AI regulation. In one treatment, we informed them that one of the videos was a deepfake and our measure is which video they elected to watch. In Study 3, participants viewed both videos (in favor of and opposed to the need for AI regulation). One of the videos was real and the other was a deepfake, which was either unlabeled or labeled. Our main outcome measure is which of the videos they shared with another participant. Finally, in Study 3b, we asked participants to watch the shared videos and asked them to evaluate the videos as well as their perceptions of the participant who shared them, Across our studies, we find that labels are effective at the extensive margin—reducing voluntary exposure and the willingness to share—but not at the intensive margin—persuasion conditional on exposure. Sharing a labeled as opposed to an unlabeled deepfake video also did not change perceptions of the sender, but people thought a labeled deepfake video was less deceptive. These results suggest that labels may help prevent the spread of false beliefs when people can opt in or out of viewing and sharing content, but are unlikely to reduce persuasion when labeled deepfakes are aired within news programming or otherwise passively encountered.

Experiments

We generated two scripts on the topic of AI regulation using ChatGPT, with one arguing in favor of AI regulation, while the other is opposed. The scripts are of approximately equal length (see the OSF repository for the full texts). We then recorded a professor reading each script with the help of a teleprompter, with each video lasting a little over four minutes. Then, we used publicly available recordings of the same professor to train an algorithm to generate two deepfake videos using LatentSync. Specifically, we took the video arguing in favor of AI regulation and generated a deepfake such that the speaker instead delivered the speech opposed to AI regulation, and vice versa ("deepfake" videos). Thus, we have for each argument (for and against regulation) a real video and a paired AI generated deepfake with the same content. We then generated a "labeled deepfake" by adding text identifying the video as "AI Generated" (see Figure 1). We used these materials throughout our experiments.

Figure 1

Screenshots displaying the real video (left side) and the watermark labeling used in the deepfake video (right side).



(A) *Real Video*

(B) *Labeled Deepfake Video*

In Study 1, we examine if adding a disclosure label for AI generated deepfake videos reduces persuasion among viewers. We recruited participants from Prolific (n = 1,800; M_{Age} = 42.5, 47.8% Male) and first measured their attitudes toward AI regulation using a six-item index, such as "Government regulation is necessary to ensure that generative artificial intelligence is used ethically," and "Government regulation will stifle innovation in generative artificial intelligence (reverse coded)" using sliders from -100 to 100, with endpoints labeled as "Strongly disagree" and "Strongly agree" and the midpoint labeled "Neither agree nor disagree." We averaged the responses across the items (three items reverse-coded, $\alpha = 0.86$).[3] We then randomly assigned them to one of six treatments, varying both the version of the video and the content of the speech. Participants viewed either the real recording of the argument, a deepfake in which the speaker was altered to present the script for the other position, or a labeled deepfake in which they were informed in advance that the video was created using artificial intelligence and the video contained a watermark. They were further randomly assigned to view either arguments for or against regulation.

After watching the video, we asked participants the same six items to measure their position on AI regulation, with their previous responses pre-selected. We calculated a measure of persuasion that is the change in the index, signed in the direction of the argument in the video. We further asked participants how persuasive they thought the video was in shaping their views (5-point Likert scale ranging from "not at all" to "extremely") and whether they thought it was generated using AI ("Yes" or "No"). We incentivized participants to pay attention by informing them prior to watching the video that they would later be asked three questions about the content and could earn a bonus of 10 cents for each question they answer correctly (on average, they answered 1.97 questions correct). The survey ended with basic demographic questions.

Our preregistered primary outcome is the attitude change in the direction of the argument in the video. We fit an OLS regression with dummy variables for whether the video is a Deepfake and whether there is a Label present, such that the coefficient for the Label estimates the impact of the disclosure on the persuasiveness of a deepfake video. We also include a dummy variable to capture whether the video advocated for or against regulation because arguments may not be

[3] For additional details on this and the following experiments, see the extended methods. Screenshots of all materials, response data, code to reproduce all figures and analyses, the videos with the real and deepfake arguments for and against AI regulation, and the AsPredicted preregistration reports are all available via OSF: https://osf.io/jfkc8/overview?view_only=03c99b09c519413b8626210f0c5d1251

equally persuasive or participants who favor or oppose regulation may not be equally persuadable. We show these results in Table 1 and, visually, in Figure 2. Notably the real video opposing regulation was very persuasive, changing beliefs by almost 18 points on average. The deepfake video was not significantly less persuasive. And, consistent with our preregistered prediction, our key finding shows that disclosure via adding a label did not reduce the impact on beliefs relative to the unlabeled deepfake. This shows that it is the content of the arguments more so than whether the source indeed made those arguments that is driving persuasion. When looking at our Likert measure (column 2), participants who were assigned to watch the deepfake video reported it to be less persuasive, but no less so if they saw the video that was labeled. This suggests that perceptions about persuasiveness may differ from actual persuasiveness. Even so, importantly, the label affects neither.

Among participants watching the labeled deepfake video, 90.4% reported that they thought it was generated with AI, which was significantly greater than those watching the unlabeled deepfake (77.8%; $\chi^2(1, n = 1198) = 34.51$, $p < .001$). Thus, the watermark informed some people who would otherwise not have noticed it. Notably, some participants did not trust the label and still believed it to be a real video. Conversely, among participants who watched the real video, 27.2% incorrectly thought it was AI generated.

Both real and deepfake videos shifted attitudes. Adding an on-screen “AI generated” label to deepfakes increased the share of participants who recognized the video as a deepfake, but, consistent with our prediction and prior work on the limits of disclosure, did not significantly reduce the video’s impact on beliefs. However, labels may still meaningfully mitigate downstream influence by reducing exposure and distribution in contexts where audiences opt in to viewing content, or have control over whether to forward it. That is, people may not wish to be exposed (or expose others) to content they know is fake. Conversely, participants may be *curious* about AI generated videos, particularly given their novelty, and thus seek them out even more.

Study 2 examines whether people avoid content that is flagged as a deepfake video. We explore this in the context of counter-attitudinal videos because persuasion campaigns are likely to target those who do not already hold the advocated view. Moreover, people have been shown to prefer content that aligns with their own position and avoid information that contradicts their beliefs (Golman et al., 2017; Hart et al., 2009). Therefore, they may avoid deepfakes expressing an opposing view regardless of whether it's labeled, rendering the label ineffective. Alternatively, they may be curious to see the deepfake and thus the label may backfire and increase exposure.

We recruited participants from Prolific (n = 1,551; M_{Age} = 45.04, 47.7% Male) and first measured their attitudes toward AI regulation using the same six-item index as in Study 1 (α = 0.8) and classified them as being either for regulation (a score above zero) or against regulation (below zero).[4] Participants then chose one of two videos to watch: one aligned with their initial attitude (the Real video) and one opposed to it (the Deepfake). Participants were randomly assigned to a Label condition in which the belief-opposing video was disclosed as "AI generated" prior to their choice and later with an embedded watermark, or a Baseline condition in which they were not informed that it was a deepfake. After viewing the chosen video, participants again reported their attitudes on the same scale. The key outcomes were (i) whether participants chose the belief-opposing video (exposure), and (ii) the magnitude of attitude change toward the opposing view, regardless of which video they chose. Participants again reported whether they thought the video was AI generated, and we incentivized them to listen to the arguments with a bonus for correctly answering questions about the content of the video.

We first estimate how labels affect people's willingness to select the belief-opposing video. Contrary to our expectation, curiosity did not lead people to opt into watching the AI generated video. Instead, the label effectively reduced exposure from 38.7% to 30.5% ($\chi^2(1, n = 1551) = 11.26$, $p < .001$). We report in column 3 of Table 1 a linear probability model that also controls the direction of the argument presented in the deepfake video. Participants who were opposed to regulation (and hence whose deepfake video corresponded to a pro-regulation argument) were more likely to select the deepfake video ($p < 0.001$). Notably, however, even controlling for this, the effect of the label still reduced willingness to select the deepfake video by

[4] 14 participants had an average index score of zero and were randomly assigned to one of the two positions.

6.3 percentage points ($p < 0.01$).

We next look at whether labeling had an impact on aggregate beliefs. For example, it may be that people who are deterred from watching the video would not have been persuaded by the argument. We therefore compare the change in beliefs across all participants in the two treatments, regardless of which video they watched. Participants who were informed that the video was AI generated prior to their choice revised their attitude by 3.2 points in the direction of their initial position—that is, they became more convinced about their initial view. Those in the Baseline condition revised their attitudes by 0.98 away from their initial position ($t(1549) = 3.07$, $p = .002$). Labeling the deepfake thus reduced the persuasive impact of the AI generated argument by deterring people from watching it who would otherwise have been persuaded. Contrary to our prediction and concern, curiosity about AI generated content did not increase voluntary exposure.

Study 3 evaluates whether disclosure and labeling reduce downstream diffusion by lowering willingness to share deepfakes, as those spreading false information may worry about reputational concerns (Altay et al., 2022). Participants recruited from Prolific (n = 1,832; M_{Age} = 44.13, 45.3% Male) began by completing the six-item scale measuring their attitude toward AI regulation (α = 0.86). Participants with a score above zero were classified as pro-regulation, and those with a score below zero as anti-regulation.[5] They then watched two videos: a real video arguing against their position, and a deepfake video with arguments consistent with their own view. Unlike in the previous study, we set the belief-congruent video to be the deepfake because participants were likely most motivated to share content that supported their own position, and we wanted to test whether knowing that this content was AI generated was sufficient to dampen that sharing motive. We then randomly assigned them to one of three conditions that differed only in the information they received about the deepfake video. In the Baseline condition, they were not informed that the video was AI generated. In the Label condition, participants were informed that the video aligned with their view was AI generated and the video had a watermark identifying it

[5] 22 participants had an average index score of zero and were randomly assigned to one of the two positions.

as such. Finally, in the Disclosure condition, participants were told that the video was a deepfake but there was no watermark. Participants then watched both videos and chose which video to share with another user. Importantly, in the Disclosure condition, participants knew that the recipient would not be informed that the video they shared was a deepfake.

Participants whose belief-congruent deepfake video was not disclosed as being AI generated shared it more than half the time (62%, $t(613) = 6.15$, $p < .001$). When participants were informed that the video they were about to watch was a deepfake, their likelihood of sharing it declined by 16.5 percentage points ($t(1219) = 5.83$, $p < .001$), and by 17.6 percentage points if they were informed and the video had a watermark that would also inform the recipient ($t(1223) = 6.24$, $p < .001$). Notably, whether the watermark was included (and hence whether the recipient would also be aware that the participant knowingly shared a deepfake video) did not reduce willingness to share significantly ($t(1216) = 0.39$, $p = .696$).

After watching both videos and making their sharing decisions, participants evaluated whether they thought each of the two videos was real or AI generated. Both disclosure and disclosure with a watermark helped participants identify that deepfake content was AI generated (83.2% and 84.8%, respectively, versus 75.1% without disclosure, both $p < 0.001$).

Accuracy did not differ with or without the watermark ($t(1216) = 0.75$, $p = .452$). Notably, we also find evidence for an implied truth effect (Pennycook et al., 2020), albeit one that here is not an error. In the absence of a label, 38.4% of participants thought the real video was in fact AI generated as well. The disclosure and watermark treatments lower this share to 24.4% and 22.1%, respectively (both $p < 0.001$). That is, about 15% of participants who were informed that one of the videos was AI generated inferred from the lack of disclosure about the other video that it must have been real.

Finally, we can examine one pathway through which disclosure may deter sharing: it can eliminate plausible deniability. People may believe that a video is a deepfake, but given some uncertainty, they nonetheless share the video to propagate their own position. Disclosure without a watermark removes this channel, even though the recipient would not be aware that the sender knowingly shared misinformation. The watermark, finally, removes the uncertainty also for the recipient, who would then know that the sender knowingly shared fake content.

To examine whether disclosure deters sharing by eliminating plausible deniability, we look at the subset of participants across the three treatments who correctly identified the deepfake video as such (i.e., not participants who may have genuinely thought that the deepfake video was real). This analysis was not preregistered. Among participants who were not informed and hence correctly inferred this on their own, 58.1% nonetheless shared the deepfake video. Among those who were informed (but who knew that the recipient would not be informed), this dropped by 13.1 percentage points ($t(964) = 4.13$, $p < .001$), and by 14.9 percentage points among those whose video was also watermarked ($t(977) = 4.70$, $p < .001$); but again, the difference between the two informed treatments was not different ($t(1021) = 0.55$, $p = .583$). This suggests that the decrease in sharing results from the reduction in plausible deniability: when people merely suspect that it was AI generated, they were nonetheless willing to share it. But when ambiguity was removed, they no longer shared the AI generated video even when recipients would remain unaware.

In order to avoid deception, we then recruited a separate sample of Prolific participants to watch the video shared by participants in Study 3 (N = 1,924; M_{Age} = 42.34, 43.6% Male). We refer to this as Study 3B. Participants were informed that the video they were going to watch was shared with them by another Prolific participant who chose between sending one of two videos, with one in favor and one opposed to AI regulation. They were not informed that the choice was between a real and a deepfake video. After watching the full video, participants answered three questions related to trust (in the speaker, the argument presented, and the participant who shared the video), a question about the perceived strength of the argument in the video, and a question about whether they would rely on the sender for other information (all on five point Likert scales ranging from “Not at all” to “Extremely”). We then asked participants if they thought the video they watched was real or AI generated, and asked them to rate their support for mandating the labeling of AI generated content on a seven point Likert scale (ranging from “Strongly oppose” to “Strongly support”). We report the results descriptively, but made no preregistered predictions.

Notably, participants were not assigned to all conditions equally because assignment depended on the videos that were shared by other participants. Because most senders (and receivers) were in favor of regulation, more real videos opposed regulation and were thus counter-attitudinal to the receivers. Because trust and agreement are correlated (Hagmann et al., 2024), we report results separately for whether the participant agreed or disagreed with the argument in the video. Figure 3 shows that participants trusted the speaker in the video, the argument, and the sender less when they watched a deepfake than a real video. They also thought the argument was weaker and were less willing to rely on the sender for information. Notably, however, this did not differ across labeled or unlabeled conditions, regardless of whether the argument was aligned or opposed to the participant's own view.

84.4% of participants correctly identified the unlabeled deepfake video as AI generated, and this increased to 91.7% with the AI Generated watermark ($t(1010) = 3.09$, $p = .002$). Notably, the lowest accuracy was for the real video: only 67% correctly identified it as a real video. Regardless of how they identified the videos, participants found the real video less deceptive than the unlabeled deepfake (1.88 vs. 2.37, $t(1633) = 8.38$, $p < .001$) and also the labeled deepfake (2.08, $t(1199) = 2.78$, $p = .005$). Notably, adding the label led to the video being perceived as less deceptive ($t(1010) = 3.27$, $p = .001$). Finally, participants expressed very high support for a mandate to label AI generated content: on a 7-point Likert scale ranging from "Strongly oppose" to "Strongly support," 69% expressed the strongest support and only 2 were opposed (average 6.46).

Discussion

Widespread availability of tools that can create realistic, but false, video content is raising concerns about the impact of such content on public opinion. People may believe that events that never happened are real, and may generally become more skeptical of news footage even when it is real. One approach to addressing this concern is to mandate labeling via text embedded in the video informing viewers of the artificial nature of the content. We show that this labeling may be partially effective: people are less likely to watch and share content that they know is AI

Table 1

OLS regression results from all three studies examining the effect of labeling on persuasion, voluntary exposure, and willingness to share deepfake content. Column 1 shows the change in beliefs about the need for AI regulation in the direction of the position advocated in the video. Column 2 shows how persuasive participants thought the video was. Column 3 shows the likelihood of watching the belief-opposing deepfake video, and Column 4 shows the degree of persuasion away from the originally held position. Finally, Column 5 shows the likelihood of sharing a belief-aligned deepfake video.

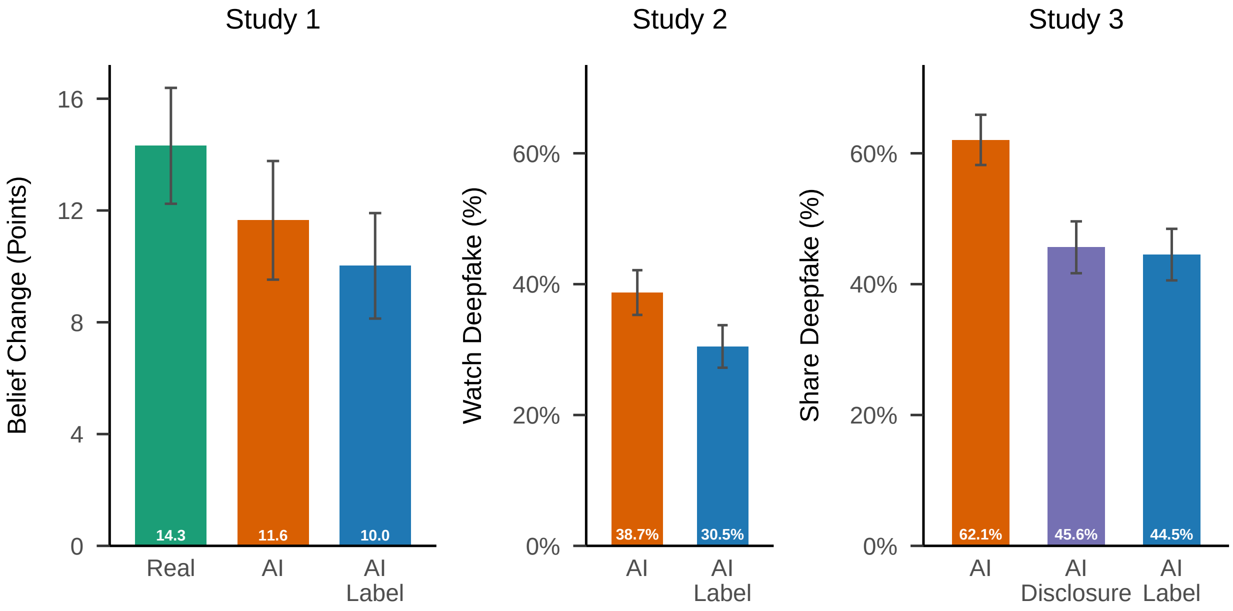
	Study 1		Study 2		Study 3
	Belief Change	Persuasiveness	Exposure	Belief Change	Share Deepfake
Deepfake (vs. Real)	−2.714+	−0.319***			
	(1.456)	(0.069)			
Label Shown	−1.542	−0.095	−0.062**	−4.178**	−0.175***
	(1.450)	(0.068)	(0.023)	(1.360)	(0.028)
Disclosure Only					−0.164***
					(0.028)
Pro-Regulation Video	−6.461***	0.324***	0.268***		
	(1.186)	(0.056)	(0.027)		
Constant	17.791***	3.040***	0.310***	0.981	0.621***
	(1.205)	(0.057)	(0.018)	(0.959)	(0.020)
N	1800	1800	1551	1551	1832

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

generated. However, we also show that this labeling has no effect on the persuasiveness of a message among those who do view it. That is, people’s opinions about a policy-relevant belief (here: support or opposition to government regulation of AI) are impacted by deepfake videos, regardless of whether they are labeled as such. This suggests that labels may be effective in contexts such as social media, in which people can choose what content to view and share, but not

Figure 2

Main results from three experiments on disclosure of AI generated content. Adding a disclosure label did not reduce the persuasiveness of the video's arguments relative to an unlabeled deepfake (Study 1). Participants were less likely to watch a belief-opposing video when they were informed that it was a deepfake (Study 2). Finally, informing participants that a belief-aligned video was a deepfake reduced their willingness to share it, even if the recipients would not be informed (Study 3). Error bars show 95% confidence intervals.



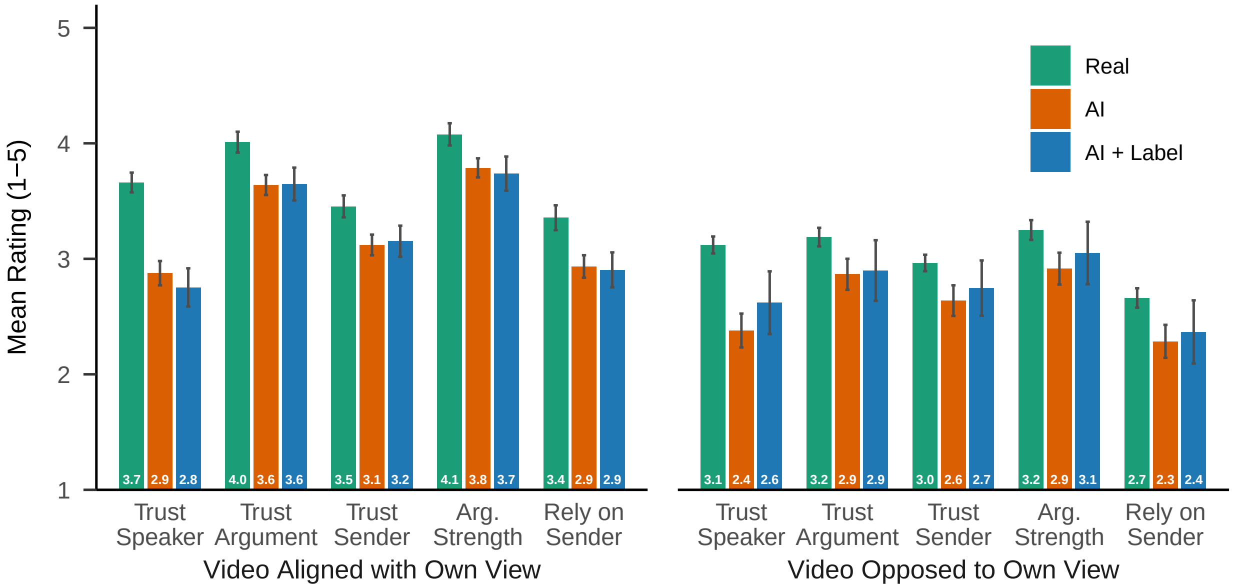
in settings where people are passively exposed to content, such as streaming news channels.

Another limitation of disclosure is that it requires cooperation from the content creator. In organized disinformation campaigns, for example, an actor seeking to deceive viewers (e.g., to influence the outcome of an election) would not abide by a disclosure requirement. While regulation could mandate embedded watermarks by video generation tools, this would limit their use in legitimate content creation (e.g., for the purpose of entertainment or education) and would, moreover, be easily circumvented by relying on open source tools.

Generative artificial intelligence is likely to have a growing impact on belief formation and public opinion. For our research, we used ChatGPT to write persuasive arguments for and against government regulation of AI. These arguments were persuasive and impacted people's views.

Figure 3

Evaluations of speaker, arguments, and sender as rated by participants in Study 3B. Deepfakes are rated as less trustworthy across dimensions, and participants are less willing to rely on the sender for other information than after watching a real video. Notably, however, we observe no difference between the unlabeled and labeled deepfake videos.



Moreover, our video format gave participants plenty of opportunities to recognize it as a deepfake: because they featured a single speaker for four minutes, with today's technology, it was not possible to create flawless deepfakes. In the future, advances in the technology could make it more difficult or impossible to detect them. Even so, many participants thought the video was real—and perhaps also concerningly, many participants thought the real video was a deepfake. Notably, this is not the only way in which AI can influence public opinion. As people increasingly chat with large language models and use them to get advice and as tutors, actors could strategically insert content that shifts the model's views on a topic—a bias that would then be passed on to the users. Moreover, custom instructions provided to models could more directly influence how they respond to questions. As far as we know, there are no regulations guiding such custom instructions, nor a requirement that they be disclosed.

Extended Methods

The studies in this paper have been approved by the human subjects review board at the Hong Kong University of Science and Technology.

Study 1. Participants first answered six questions related to their views on AI regulation: “Government regulation is necessary to ensure that generative artificial intelligence is used ethically,” “I believe that without government oversight, generative artificial intelligence could be harmful to society,” “The government should impose strict regulations on the development and deployment of generative artificial intelligence,” “Government regulation will stifle innovation in generative artificial intelligence,” “Government interference in generative artificial intelligence should be minimal to allow for technological advancement,” and “Regulation of generative artificial intelligence could hinder its potential positive impact.” All questions were asked on sliders ranging from -100 (Strongly disagree) to +100 (Strongly agree), with a neutral 0 point labeled as “Neither agree nor disagree.” To calculate our measure of support for AI regulation, we reverse-coded the last three items and averaged participants’ responses.

After the baseline attitude measure, participants were informed that they would watch a video in which a professor of AI discussed whether the use of artificial intelligence should be regulated by the government. They were informed that the video was approximately four minutes long, and that they would be able to earn a bonus of up to 30 cents for answering three multiple-choice questions about the content of the video.

We created six videos presenting arguments for and against AI regulation as discussed in the main section of the paper. Participants were randomly assigned to watch one of the videos in a 3 × 2 design. On one dimension, we varied whether the video they saw was Real, a Deepfake, or a Labeled Deepfake (which included a watermark showing “AI Generated”); and on the other dimension, the video either made arguments in favor of or against government regulation of generative artificial intelligence. In the Labeled Deepfake treatment, participants were additionally informed prior to watching the video that while the professor was real, the video itself was created by artificial intelligence.

After watching the assigned video, participants were then presented with the identical six-item scale from the beginning of the survey, with the sliders placed at the locations of their initial response. We again calculated an average score after reverse coding three of the items, and our main measure of belief change was the difference in the scale in the direction of the argument made in the video (e.g., *Posterior_Score* - *Prior_Score* for those viewing an argument in favor of regulation). We then asked them "how persuasive do you think the video was in shaping your views on artificial intelligence regulation," on a five-point Likert scale ranging from "Not at all persuasive" to "Extremely persuasive."

Participants then answered three multiple-choice questions, each with three options, about the content of the video. The questions were identical for the Real, Deepfake, and Labeled Deepfake conditions, but differed for videos in favor of regulation and those opposed to regulation, as those videos contained different arguments. Participants were then asked if they thought the video was in fact generated using artificial intelligence (yes or no), and the survey concluded with basic demographic questions (age, education, gender, ethnicity, race, and political affiliation) and an open-ended box for comments for the researchers.

Repeat Participants. The study we report here is a replication of a design in which we asked the six-item scale only after watching the video. We collected the data presented here after running Study 2. Due to an error setting the eligibility criteria on Prolific, 351 participants who had previously completed a related study were able to enter again. Because they might have been assigned to different treatments previously, we excluded all duplicate respondents and increased our sample to arrive at the preregistered number of first-time respondents. The full data with duplicates is available on OSF and all results replicate with the full sample.

Study 2. Participants first reported their beliefs on AI regulation using the identical six item scale as in Study 1. Based on their responses, we classified them as Pro Regulation (an average above zero) or Anti Regulation (below zero), with those who scored exactly zero randomly allocated to one side. They were then given the choice to watch one of two videos: one explaining why artificial intelligence should be regulated, or one explaining why it should **not** be regulated. For all participants, the video opposing their initial view was the deepfake video. In the Baseline condition, participants were not aware of this when they made their choice and they saw no indication that the video was AI generated. In the Label condition, participants were informed which of the two videos was a deepfake. We also disclosed how we generated the video: by taking the recording arguing for the opposite position and using specialized software to replace the words.

Our main outcome measure was which of the two videos they elected to watch: the real, belief-congruent video or the deepfake that opposed their initial view. After making their choice, we informed participants that they would earn a bonus for answering three questions about the content of the videos at the end of the survey to incentivize them to pay attention. They then watched their chosen video, before again answering the six-item scale of support for AI regulation with their initial responses selected by default. We calculate the persuasiveness of the videos, conditional on exposure, by looking at the change in the scale signed in the direction of the argument in the video. In addition, we looked at the extent to which a deepfake "campaign" would have been effective, examining persuasion in the direction of the fake video regardless of the video the participants elected to watch.

The survey then concluded in the same way as Study 1: participants answered the identical three content questions, whether they thought the video they had watched was generated by artificial intelligence, provided demographic information, as well as optional comments for the researchers.

Study 3. Paralleling our previous two studies, participants began by completing the six-item scale on AI regulation attitude. As in Study 2, we assigned participants to be Pro (Anti) Regulation if their average response was above (below) zero, and randomly assigned participants to one side if their score was precisely zero. We then informed them that they would watch two videos on AI regulation, one making an argument in favor and one making an argument against more regulation. For all participants, the video aligned with their initial position was the deepfake.

In the "Baseline" condition, participants got no further information. In the Disclosure and

Label conditions, participants were informed that the two videos further differed in how they were created: one video was a real recording of a professor, while the other video was created by replacing the text in the video. In the Label condition, the video also contained watermark text identifying it as “AI Generated.” We informed them that they would be asked three questions about the content of the two videos, and participants watched the two videos in random order. They then selected one of the two videos they wanted to share with a participant we would subsequently recruit. Participants saw a screenshot from each of the two videos and selected whether to share the one in favor or the one against regulation. In the disclosure and label conditions, they were again reminded that the video aligned with their position was a deepfake, and in the label condition the screenshot also showed the “AI Generated” watermark.

Finally, participants answered the three content questions, indicated whether they thought each of the two videos was real or AI generated, and answered the identical demographic questions and optional open-ended comment as in the previous two studies.

Study 3B. We then recruited a new sample of participants to watch the videos that had been selected for sharing. All participants first reported their attitudes toward AI regulation using the same six item scale as in the previous studies. They were informed that another Prolific participant had previously watched two videos, one arguing in favor of AI regulation and one arguing against it, and had chosen to share one of them. Depending on the condition and the decision of the original participant, the shared video was either a real recording, a deepfake without a label, or a deepfake with a label. After viewing the video, participants rated the trustworthiness of the speaker, the argument, and the person who shared the video, as well as the perceived strength of the argument and how much they would rely on the sender for other information. Participants also reported how deceptive they found the video. All measures were assessed on five-point scales. Finally, participants indicated whether they believed the video was real or AI generated and reported their support for requiring labeling AI generated content on a seven point scale ranging from “Strongly oppose” to “Strongly support”.

References

Allen, J., Watts, D. J., & Rand, D. G. (2024). Quantifying the impact of misinformation and vaccine-skeptical content on Facebook. *Science*, *384*(6699), eadk3451. https://doi.org/10.1126/science.adk3451

Altay, S., Hacquin, A.-S., & Mercier, H. (2022). Why do so few people share fake news? It hurts their reputation. *New Media & Society*, *24*(6), 1303–1324. https://doi.org/10.1177/1461444820969893

Ben-Shahar, O., & Schneider, C. E. (2014). *More Than You Wanted to Know: The Failure of Mandated Disclosure*. Princeton University Press. https://doi.org/10.2307/j.ctt5hhrqj

Berlyne, D. E. (1960). *Conflict, arousal, and curiosity* (pp. xii, 350). McGraw-Hill Book Company. https://doi.org/10.1037/11164-000

Block, M. J. (2024). A Critical Evaluation of Deepfake Regulation through the AI Act in the European Union. *Journal of European Consumer and Market Law*, *13*(4), 184–192.

Cain, D. M., Loewenstein, G., & Moore, D. A. (2005). The Dirt on Coming Clean: Perverse Effects of Disclosing Conflicts of Interest. *The Journal of Legal Studies*, *34*(1), 1–25. https://doi.org/10.1086/426699

Carney, S., Riveros, I., & Tully, S. (2024). *Made with AI: Consumer Engagement with Media Containing AI Disclosures*. SSRN. https://doi.org/10.2139/ssrn.4988760

Carrella, F., Simchon, A., Edwards, M., & Lewandowsky, S. (2025). Warning people that they are being microtargeted fails to eliminate persuasive advantage. *Communications Psychology*, *3*(1), 15. https://doi.org/10.1038/s44271-025-00188-8

Chesney, B., & Citron, D. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, *107*(6), 1753–1820.

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*(6714), eadq1814. https://doi.org/10.1126/science.adq1814

Gallegos, I. O., Shani, C., Shi, W., Bianchi, F., Gainsburg, I., Jurafsky, D., & Willer, R. (2025).

Labeling Messages as AI-Generated Does Not Reduce Their Persuasive Effects (arXiv:2504.09865). arXiv. https://doi.org/10.48550/arXiv.2504.09865

Golman, R., Hagmann, D., & Loewenstein, G. (2017). Information Avoidance. *Journal of Economic Literature*, *55*(1), 96–135. https://doi.org/10.1257/jel.20151245

Green, M., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology*, *79*(5), 701–721. https://doi.org/10.1037/0022-3514.79.5.701

Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025). The levers of political persuasion with conversational artificial intelligence. *Science*, *390*(6777), eaea3884. https://doi.org/10.1126/science.aea3884

Hagmann, D., Minson, J. A., & Tinsley, C. H. (2024). Personal narratives build trust across ideological divides. *Journal of Applied Psychology*, *109*(11), 1693–1715. https://doi.org/10.1037/apl0001201

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, *135*(4), 555–588. https://doi.org/10.1037/a0015701

La Ferrara, E., Chong, A., & Duryea, S. (2012). Soap Operas and Fertility: Evidence from Brazil. *American Economic Journal: Applied Economics*, *4*(4), 1–31. https://doi.org/10.1257/app.4.4.1

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. https://doi.org/10.1177/1529100612451018

Lin, H., Czarnek, G., Lewis, B., White, J. P., Berinsky, A. J., Costello, T., Pennycook, G., & Rand, D. G. (2025). Persuading voters using human–artificial intelligence dialogues. *Nature*, *648*(8093), 394–401. https://doi.org/10.1038/s41586-025-09771-9

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, *116*(1), 75–98. https://doi.org/10.1037/0033-2909.116.1.75

Loewenstein, G., Sunstein, C. R., & Golman, R. (2014). Disclosure: Psychology Changes Everything. *Annual Review of Economics*, *6*(1), 391–419. https://doi.org/10.1146/annurev-economics-080213-041341

Lu, L., Tormala, Z. L., & Duhachek, A. (2025). How AI sources can increase openness to opposing views. *Scientific Reports*, *15*(1), 17170. https://doi.org/10.1038/s41598-025-00791-z

Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, *119*(8), e2120481119. https://doi.org/10.1073/pnas.2120481119

Paris, B., & Donovan, J. (2019). *Deepfakes and Cheap Fakes*.

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings. *Management Science*, *66*(11), 4944–4957. https://doi.org/10.1287/mnsc.2019.3478

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595. https://doi.org/10.1038/s41586-021-03344-2

Sah, S., Loewenstein, G., & Cain, D. M. (2013). The burden of disclosure: Increased compliance with distrusted advice. *Journal of Personality and Social Psychology*, *104*(2), 289–304. https://doi.org/10.1037/a0030527

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*, *6*(1), 2056305120903408. https://doi.org/10.1177/2056305120903408

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Weil, D., Graham, M., & Fung, A. (2013). Targeting Transparency. *Science*, *340*(6139), 1410–1411. https://doi.org/10.1126/science.1233480

Wittenberg, C., Epstein, Z., Péloquin-Skulski, G., Berinsky, A. J., & Rand, D. G. (2025). Labeling AI-generated media online. *PNAS Nexus*, *4*(6), pgaf170. https://doi.org/10.1093/pnasnexus/pgaf170