Deepfakes and the Erosion of U.S. Democracy: Societal Trust and AI Regulation

Natalie S.K. Ringdahl

Department of Social Sciences, University of California, Irvine

Faculty Advisors: Dr. Sara W. Goodman and Dr. Sven Bernecker

May 12, 2025

Author Note

Natalie S.K. Ringdahl, Department of Social Sciences and Humanities, University of California, Irvine. This thesis was submitted in partial fulfillment of the requirements for the Political Science Honors Program and the Campuswide Honors Collegium (CHC). This work contributes to the broader academic discourse on AI ethics, disinformation, and policy regulation, specifically analyzing how deepfake technology affects societal trust in U.S. democratic institutions and evaluating the effectiveness of AI regulation in California. Correspondence regarding this thesis should be directed to nringdah@uci.edu.

Abstract

The rise of AI-generated deepfakes presents a crucial challenge to societal trust in US governmental institutions by undermining privacy and the authenticity of information. This study investigates how deepfakes impact societal trust and evaluates public perceptions of AI regulation through California's SB 1047 and SB 942. Using a randomized survey experiment with 104 participants, the study measured changes in trust, concern about disinformation, and policy perceptions following exposure to real and AI-manipulated media. Results showed strong public support for robust AI regulation, but moderate skepticism toward transparency-based solutions like content labeling. Perceived realism of deepfakes was modestly associated with trust erosion, while concerns about privacy and democratic integrity remained high across all participants. These findings highlight the need for proactive regulatory frameworks and civic education efforts to safeguard democratic trust in an era of easily fabricated media and uncertain epistemic agency.

Deepfakes, Societal Trust, and AI Regulation in U.S. Democracy

Perhaps the most beautiful facet of AI is also its most concerning: it is intangible yet omnipresent, a borderless inevitability permeating human life. Without clear regulation, AI media manipulation can undermine the very essence of identity and trust in society, thus posing a significant threat to democracy. Currently, AI is capable of recreating and manipulating a person's appearance, voice, and likeness – a phenomenon known as 'deepfake.' These AI-generated fake images, videos, and audio recordings pose a global threat to the democratic right of privacy and authenticity. Such threats can only be mitigated through thoughtful regulation and ethical considerations.

Congress has already acknowledged the importance of mitigating these highly damaging and often non-consensual deepfakes in the bipartisan passing of the "Take It Down" Act on April 28, 2025. The unanimous passing of this bill by the Senate makes it illegal to "knowingly publish" or threaten to publish intimate images without a person's consent, requiring websites and social media companies to remove such material within 48 hours of notice from a victim. Most notably, this bill departs from previous legislation banning similar sexually explicit deepfake and "revenge porn" content because federal regulators are directly imposing on internet companies. The overwhelming support from Congress and the First Lady Melania Trump not only demonstrates the intensity of national concern and scrutiny, but reveals an inherent skepticism toward the ethical use of AI-manipulated media.

This paper first examines existing literature on trust, disinformation, and AI regulation, then presents experimental findings, followed by a discussion on policy effectiveness. It is guided by two central research questions:

1) How do deepfakes weaken societal trust in U.S. governmental institutions by challenging authenticity and privacy?
2) How effective are California's legislative efforts—particularly SB 1047 and SB 942—in mitigating the risks posed by AI technologies like deepfakes?

To address these questions, the study investigates two dependent variables: (1) levels of societal trust in government institutions, and (2) public perceptions of AI regulatory effectiveness. Two hypotheses structure the analysis: first, that stronger perceived effectiveness of AI regulatory principles (such as transparency and privacy protections) will correlate with higher societal trust; second, that exposure to deepfake media will negatively affect participants' trust in U.S. governmental institutions.

In light of these findings, the paper concludes with policy recommendations aimed at improving regulatory responses to deepfake media, particularly through public perception of California's AI legislation.

Findings from a randomized survey experiment with 104 participants indicate strong public concern about privacy and disinformation, widespread support for robust AI regulation, and mixed confidence in transparency-based solutions like content labeling. Participants exposed to deepfakes showed slightly decreased trust in public figures and U.S. governmental institutions—particularly when the manipulated media appeared more realistic—though most could detect the deepfake. Importantly, even those who were not deceived expressed deep concern about the implications of such technology for democratic integrity.

In an age where truth itself can be fabricated, this research underscores the urgency of proactive regulation, civic education, and institutional transparency. California's evolving AI legislation offers both a case study and a cautionary tale: without meaningful public trust, even

well-intended policies may fall short. Deepfakes threaten not only what we believe, but how we come to believe it, thus making the defense of epistemic agency essential for preserving democracy.

Background

The rise of deepfake technology has elevated long-standing concerns about disinformation into a new realm of realism and reach. Unlike earlier forms of misinformation that could often be traced, contextualized, or corrected, deepfakes blur the boundary between what is real and what is fabricated—posing unprecedented challenges for truth, privacy, and institutional trust. Unlike traditional misinformation, the realism of deepfakes makes them harder to detect, easier to believe, and more damaging to public perception.

This growing epistemic instability is not only a technical challenge, but a governance dilemma. Legal scholars and policy analysts have noted that traditional regulatory tools—such as content takedown laws or platform-specific guidelines—struggle to address the distributed and evolving nature of AI-generated media. Transparency-based measures, like labeling content as AI-generated, have been proposed and even enacted in states like California. Yet critics argue these policies may be insufficient, especially when bad actors operate across borders, algorithms outpace detection tools, or users fail to internalize labels as meaningful cues. Meanwhile, more comprehensive regulatory efforts, such as SB 1047, have faced political resistance on the grounds that they might stifle innovation.

At the same time, deepfakes raise pressing questions about the erosion of public trust. As manipulated media becomes more sophisticated and widespread, so too does the risk that citizens begin to doubt not only political figures and media outlets, but the very possibility of objective information. Scholars like Coeckelbergh (2024) and Vaccari & Chadwick (2020) have argued

that deepfakes do not merely deceive—they destabilize the public's ability to discern truth, undermining the foundations of informed democratic participation.

In this context, understanding how deepfakes influence public trust and how policy can meaningfully intervene is critical. This study builds on these debates by investigating the psychological impact of deepfake exposure and the perceived effectiveness of transparency and accountability based regulatory frameworks in California.

Literature Review

The rise of artificial intelligence (AI) has introduced both unprecedented opportunities and profound risks, particularly in the realm of information authenticity and democratic stability. This literature review examines two major consequences of deepfake technology:

(1) its potential to erode public trust in democratic institutions, and

(2) the difficulty of regulating AI-generated media through existing legal frameworks.

The first section explores how philosophers and political theorists define trust, and how empirical studies have shown that deepfakes disrupt the conditions necessary for maintaining collective belief in government transparency, accountability, and authenticity. The second section examines legal and policy responses to deepfakes focusing especially on the transparency-and-oversight based approaches embedded in California's SB 942 and the proposed SB 1047, and evaluates scholarly debates over whether these regulatory efforts are sufficient. Together, these two threads lay the foundation for this study's hypotheses about how deepfakes influence both trust and perceptions of regulatory effectiveness.

Trust in Institutions

For the purposes of this paper, trust is understood as the relationship between a collective and an institution. Philosophers and political scholars have long examined the notion of trust and have had great difficulty in creating one uniform definition. Broadly speaking, trust has been explored as an abstract interplay between faith, legal assurances, and social expectations, forming a foundation for societal cooperation.

As Heimer (2001) argues, trust is embedded in institutional and legal structures, helping societies manage uncertainty through both formal and informal mechanisms. This literature on trust provides a backdrop for Gilbert's concept of collective belief, which frames trust in this paper as a collectively maintained expectation—one that citizens in the United States place in their government to uphold democratic values of transparency, authenticity, and accountability (Gilbert, 1987). Similarly, Devos et al. (2002) found that trust in institutions is closely linked to the values individuals associate with those institutions, shaping their level of confidence in governmental actions and policies. Beyond mere expectation, citizens' collective belief in government carries an implicit commitment that institutions will actively uphold democratic values and rights. As Hawley (2014) argues, trust in institutions extends beyond mere belief in their intentions; it reflects an expectation that they are both willing and able to fulfill their democratic duties in good faith. This aligns with Hetherington's (2005) assertion that political trust is rooted in the perception that government acts in the public interest, ensuring that transparency, authenticity, and accountability are not just ideals, but actively sustained democratic principles.

However, in an era where AI enables the subtle fabrication of misleading content, this collective belief in governmental transparency and authenticity is challenged. The rise of

deepfakes and AI-generated disinformation threatens to distort reality and erode trust in democratic institutions.

Deepfakes and Disinformation

In 2018, BuzzFeed released a viral deepfake video of former President Barack Obama, featuring manipulated speech by filmmaker Jordan Peele, to illustrate the dangers of AI-generated disinformation (BuzzFeed, 2018). The Obama deepfake was an early glimpse into the concerning powers of AI-manipulated media and its potential to distort public perceptions. Since then, deepfake technology has evolved dramatically, making it increasingly difficult for audiences to distinguish between reality and deception.

The spread of deepfakes does not simply introduce new challenges for information integrity—it directly weakens societal trust by fueling disinformation. Vaccari & Chadwick (2020) found that exposure to deepfakes induces skepticism toward news sources, showing how AI-driven deception fosters uncertainty about what is real and who can be trusted. When manipulated media circulates unchecked, it erodes the perceived authenticity of information—a key pillar of institutional trust (Heimer, 2001).

Beyond simply raising doubts about the news, deepfakes contribute to a growing culture of distrust that extends to political institutions. Coeckelbergh (2024) warns that AI-driven manipulation does not merely deceive individuals—it destabilizes collective trust by distorting the foundations of knowledge itself. As citizens become increasingly unsure of the validity of political figures' statements and are fed disinformation, they lose their epistemic agency and ability to assess truth independently, fracturing the trust that democracy depends on.

In democratic societies, trust in institutions is not an abstract ideal—it is the foundation upon which legitimacy and governance rest. Deepfake disinformation strikes at the heart of

democratic values like transparency, authenticity, and accountability. Hetherington (2005) argues that political trust is essential for democratic stability, yet deepfakes create doubt, making it easier for bad actors to manipulate elections and erode confidence in government processes. Chesney & Citron (2019) describe deepfakes as a weaponized form of disinformation, enabling these bad actors to sow chaos, fabricate political scandals, and manipulate public perception at an unprecedented rate. Similarly, Dan et al. (2021) highlights how deepfakes pose a direct threat to electoral integrity—if voters cannot trust the authenticity of political content, democratic participation itself is jeopardized.

Ultimately, deepfakes don't just blur the truth–they create a structural weakness in democracy itself. When institutional authenticity crumbles, the ability to make informed decisions vanishes, leaving democratic processes vulnerable to manipulation.

California's Approach to Deepfake and AI Regulation

The rapid advancement of AI has forced governments to confront an urgent question: how can AI be regulated effectively while preserving democratic values such as transparency and accountability? Efforts like Meta's content regulations[1] and U.S. election transparency laws[2] were designed to counter misinformation, but deepfake technology presents distinct regulatory challenges that require a more robust approach. Unlike traditional misinformation, deepfakes use synthetic media to fabricate hyper-realistic content that is false, making detection and mitigation far more complex (Wachter et al., 2021). As a result, policymakers must not only prevent the spread of deceptive AI-generated media but also restore public trust in institutions undermined by deepfake disinformation.

[1] Meta's Ad Standards outline disclosure requirements for political and social issue advertisements.
[2] The AI Transparency in Political Ads Act proposes mandatory disclosures for AI-generated political content.

California has positioned itself as a leader in AI governance, introducing legislation such as SB 942 (AI Transparency Act) and SB 1047 (Safe and Secure Innovation Act) to address the risks posed by deepfake disinformation. These policies, however, represent two contrasting approaches: SB 942 prioritizes transparency and consumer awareness by requiring AI disclosures, whereas SB 1047 proposed a comprehensive regulatory framework for AI development and deployment. While the effectiveness of these laws remains a subject of debate, critical questions arise about whether current regulatory measures can meaningfully counteract AI-driven disinformation or if they inadvertently stifle technological innovation.

California's AI Regulation: SB 942 & SB 1047

To address concerns over AI-generated disinformation, California passed SB 942 (AI Transparency Act) in August of 2024, mandating label requirements for AI-generated content. The law is designed to help users distinguish between real and manipulated media, mitigating the risk of deepfake disinformation by ensuring transparency in digital communication. By requiring clear AI disclosures, policymakers argue that SB 942 will enhance accountability for digital content creation, thereby preserving public trust in online information ecosystems (California Governor's Office, September 2024).

However, some scholars question whether transparency alone is sufficient to curb the dangers posed by deepfake disinformation. Research suggests that merely labeling AI-generated content may not effectively deter deception (Dan et al., 2021). This is because individuals with strong pre-existing biases or low media literacy may still believe and unknowingly spread deepfakes. Additionally, critics argue that bad actors can evade such transparency laws by distributing deepfake content across unregulated platforms or international domains (Chesney & Citron, 2019). Similarly, Yanamala et al. (2023) highlight the challenges of enforcing data

protection regulations, noting that while transparency measures like the General Data Protection Regulation[3] and the California Consumer Privacy Act[4] aims to increase accountability, their impact on mitigating disinformation remains uncertain.

In contrast, SB 1047 (Safe and Secure Innovation Act) proposed a comprehensive AI oversight framework, requiring developers of advanced AI models to meet specific safety and ethical standards. The bill aimed to establish accountability mechanisms for AI developers, ensuring that high-risk AI models underwent rigorous evaluation and implemented strict safety protocols before releasing their technology to the public. However, Governor Gavin Newsom vetoed SB 1047, arguing that the bill risked stifling innovation and placing unnecessary burdens on AI companies (California Governor's Office, 2024). Newsom instead advocated for a more flexible, adaptive approach to AI governance, one that supports technological progress while minimizing risk.

Balancing Innovation & Regulation: The Debate

The veto of SB 1047 underscores a fundamental debate in AI regulation: should governments prioritize strict oversight to mitigate risks, or should they foster AI innovation with minimal regulatory intervention? While California moved forward with SB 942, which requires AI-generated content labeling, the rejection of SB 1047 signals a reluctance to impose broader oversight on AI development itself. This raises concerns about whether policymakers are taking a reactive rather than proactive stance toward AI regulation.

Some social media platforms, like Meta, have implemented automated oversight for political and social issue advertisements (e.g., disclaimers about sponsors and disclosures about

[3](GDPR) is a comprehensive European Union law enacted in 2018 that governs data privacy and protection. While primarily designed to give individuals control over their personal data, it also establishes transparency and accountability obligations for organizations that process or disseminate digital information.

[4] (CCPA), enacted in 2018, is a state-level data privacy law that grants California residents new rights over their personal information, including rights to access, delete, and opt out of the sale of their data.

AI usage in media content), but these measures may be insufficient to prevent the spread of misinformation. Wachter et al. (2021) argue that AI fairness and transparency cannot be automated, meaning that regulatory interventions must extend beyond content labeling to include stricter accountability measures for AI developers; essentially, arguing for a proactive rather than reactive approach. While SB 942 introduces an enforcement mechanism by requiring AI-generated content disclosures, it remains unclear whether such transparency mandates alone are sufficient to prevent the spread of disinformation. Without stronger enforcement targeting AI misuse, deepfake disinformation may continue to erode public trust.

Unlike California's transparency-first approach, the EU model categorizes AI applications by risk level, imposing stricter compliance standards for high-risk AI models, including deepfake technologies (Binns, 2018). Rather than relying solely on transparency mandates, the EU framework integrates accountability requirements for AI developers, raising questions about whether U.S. regulations should adopt a similarly proactive strategy to mitigate deepfake disinformation before it reaches the public. Ultimately, given the growing sophistication of AI-generated disinformation, a proactive regulatory approach may be crucial to preserving societal trust and safeguarding U.S. governmental institutions.

Research Questions and Hypotheses

This study is guided by two central research questions that explore the relationship between AI media manipulation and societal trust in democratic institutions. The first question asks: *How do deepfakes weaken societal trust in U.S. governmental institutions by challenging authenticity and privacy*? Prior research suggests that individuals' perceptions of regulatory adequacy—including beliefs about transparency and oversight—are important predictors of

institutional trust (Devos, Spini, & Schwartz, 2002). In contexts where regulation is perceived as insufficient, trust in governing institutions tends to erode. Based on this, the first hypothesis is:

- H_0 (Null Hypothesis): The perceived effectiveness of AI regulatory principles has no effect on societal trust in U.S. governmental institutions.
- H_1 (Alternative Hypothesis): The perceived effectiveness of AI regulatory principles does have an effect on societal trust in U.S. governmental institutions.

To explore this, participants were exposed to either real or deepfaked media, and their post-exposure trust levels and concern ratings were measured. The second research question considers the regulatory side: *How effective are California's legislative efforts—particularly SB 1047 and SB 942—in mitigating the risks posed by AI technologies like deepfakes?* Deepfakes are a particularly disruptive form of media because they create epistemic uncertainty, destabilizing the ability to distinguish real from fake information (Coeckelbergh, 2024; Vaccari & Chadwick, 2020). This uncertainty has been shown to decrease confidence in news and media, thereby weakening broader societal trust. Accordingly, the second hypothesis is:

- H_0 (Null Hypothesis): Deepfake exposure has no effect on societal trust in U.S. governmental institutions.
- H_1 (Alternative Hypothesis): Deepfake exposure has an effect on societal trust in U.S. governmental institutions.

Together, these questions and hypotheses provide a dual lens through which to examine both the *perceptual* (how people feel about regulation) and *psychological* (how people respond to deepfake exposure) dimensions of trust in the digital age.

Method

To test the hypotheses presented, this study employed a survey-based experimental design with a deepfake exposure component. The experiment was divided into three stages: (1) pre-exposure baseline questions, (2) randomized video exposure, and (3) post-exposure evaluation. This design enabled the analysis of both perceptual attitudes toward AI regulation and psychological reactions to manipulated media.

Participants

A total of 104 participants completed the survey. Respondents were recruited through convenience sampling, with distribution via the researcher's personal, academic, and university-affiliated networks, including social media, student groups, and campus mailing lists. This recruitment strategy also generated a limited snowball effect, as some participants shared the survey with peers.

Data collection took place between March 29 and April 10, 2025. While the majority of participants were undergraduate students based in California, the sample included respondents from at least 11 other U.S. states as well as two international participants from Germany and Taiwan. Approximately 87.5% of participants reported residing primarily in California at the time of the study. Additionally, all participants were over the age of 18 and provided informed consent prior to beginning the survey.

Materials

The study was conducted using Google Forms, which allowed for automatic response collection and randomization. Participants were randomly assigned to one of two experimental conditions based on their selection of either "Hotdog" or "Hamburger"—a neutral mechanism designed to disguise group assignment.

- *Control group ("Hotdog")*: Viewed two authentic video clips of President Obama.
- *Treatment group ("Hamburger")*: Viewed one authentic video followed by a deepfake video of President Obama which used AI-generated voice and facial manipulation.

The videos were 30 seconds in length and designed to be similar in tone and delivery. Following video exposure, participants were presented with a series of Likert-scale questions evaluating trust, concern, and perceived realism.

Procedure

The questionnaire included 18 questions and was structured in three stages:

1. *Preliminary Section (Baseline Measurement)*

   Participants answered four questions measuring their baseline trust in U.S. political institutions, familiarity with deepfakes, concern about misinformation, and perceived effectiveness of current AI laws.

2. *Experimental Exposure*

   Participants were randomly assigned to a group and shown two video clips. They were then asked to identify which video seemed less authentic and which they believed was the deepfake. The realism of the second video was also rated.

3. *Post-Exposure Section*

   All participants, regardless of group, responded to five additional questions measuring changes in trust, concern about misinformation, privacy, perceived threat to democracy, and perceptions of AI regulatory principles (e.g., transparency, need for stronger regulation, trust in government action).

Ethical Considerations

This study was reviewed and approved by the University of California, Irvine Institutional Review Board (IRB) and received an exemption under protocol #6466, granted in January 2024. All participants were over the age of 18 and provided informed consent prior to beginning the survey.

To minimize bias, the AI-generated deepfake content was not disclosed until participants completed the experimental portion of the survey. All participants were fully debriefed following exposure and were given the opportunity to ask questions or withdraw their data.

The researcher also completed CITI Program training in Human Subjects Research (Social/Behavioral Investigators – Basic Course) in compliance with university research standards. Certification was completed on January 8, 2025, and is valid through January 8, 2030 (Record ID: 67158732).

This study was conducted under the supervision of a faculty research advisor and adhered to ethical standards for minimal-risk, survey-based research.

Results

This section presents the findings of the survey-based experiment designed to test two hypotheses concerning the effectiveness of AI regulation and the impact of deepfake media on institutional trust. A total of 104 participants completed the survey and were randomly assigned to either a control or treatment group. One-sample t-tests were conducted to analyze perceptions of regulatory effectiveness, and Pearson correlation coefficients were used to assess relationships between deepfake realism and post-exposure trust levels. Participants were randomly assigned to either a control or treatment group, with the latter viewing a deepfake.

Deepfake Detection Accuracy

When the control group (Hotdog) was asked which of the two videos–both authentic clips of President Obama–they believed to be a deepfake (Q5b), 48.7% selected the first clip, 20.5% selected the second, and 30.8% responded "I'm not sure." Conversely, among the treatment group (Hamburger), who were shown a manipulated deepfake clip, 84.9% correctly identified the deepfake, 5.5% incorrectly selected the authentic clip, and 9.6% responded "I'm not sure." These results suggest that while deepfakes can be detected with relatively high accuracy in controlled settings, a notable proportion of participants still expressed uncertainty.

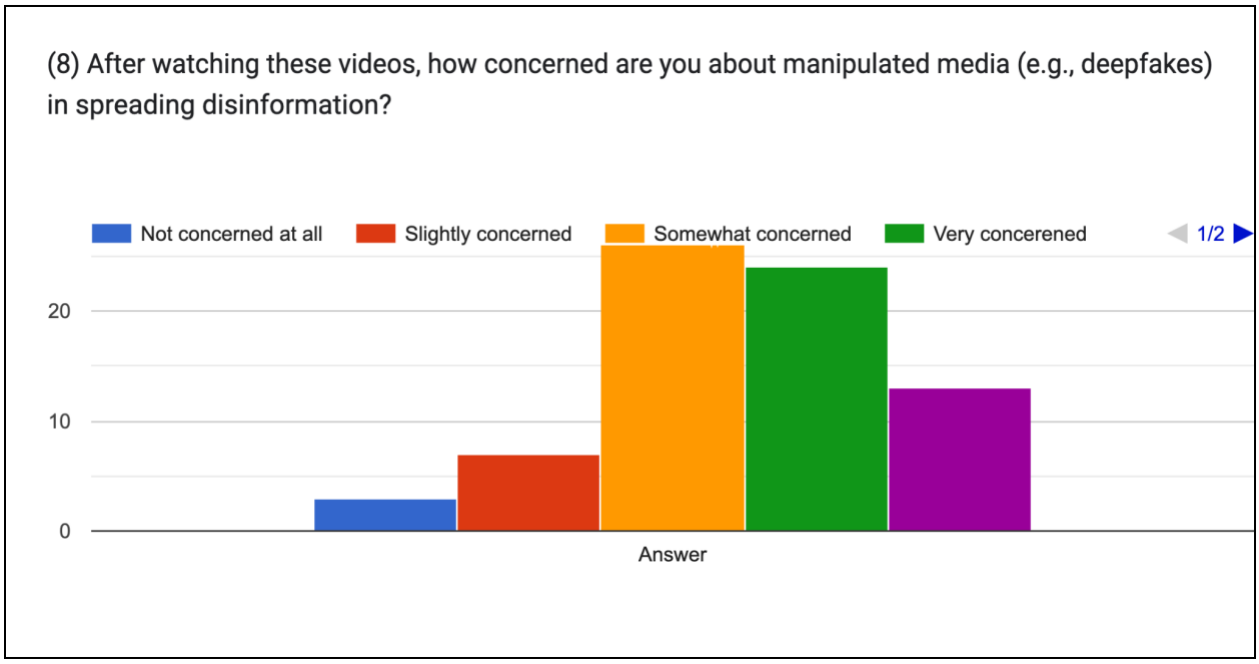
Deepfake Concerns: Privacy Infringement & Threats to Democracy

Analyses testing the relationship between perceived realism of deepfake media and trust-related outcomes (Q6 vs Q7-Q10) were limited to participants in the treatment condition, as only these respondents were exposed to manipulated content. Broader perception measures (Q11-Q15) were analyzed across the full sample.

When asked how realistic the second clip—the deepfake—appeared (Q6), 26% of treatment group participants rated it as "somewhat realistic," 34% as "slightly realistic," and 23% as "not realistic at all." Despite relatively low perceived realism, Figure 1 illustrates that 86% of these same participants reported being "somewhat" to "extremely" concerned about the spread of disinformation through manipulated media (Q8).

Figure 1

Deepfake-Induced Disinformation Concerns.



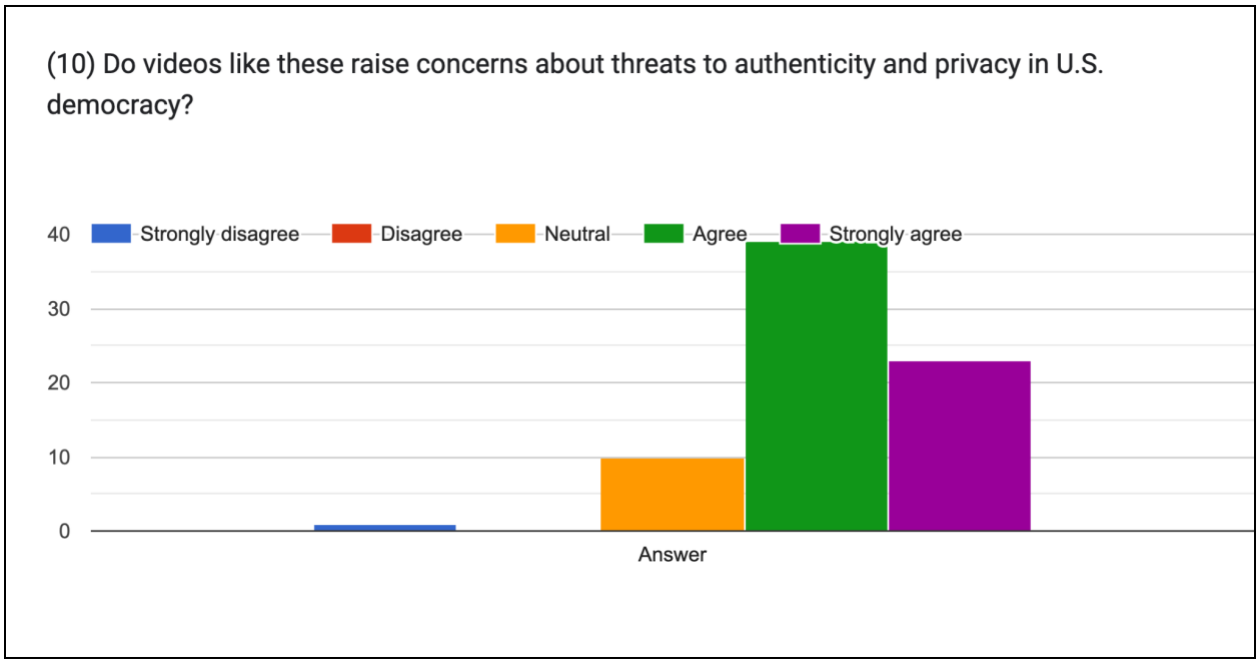
Note. The purple bar represents the proportion of respondents who selected extremely concerned.

After viewing the video clips, participants in the treatment group were asked whether the videos, including the deepfakes, increased their concerns about privacy infringement (Q9). In response, 84% of participants indicated being “somewhat” to “extremely” concerned.

Finally, concerns about threats to the authenticity and privacy in U.S. democracy were similarly high (Q10). A total of 85% of the treatment group respondents agreed or strongly agreed that the videos, including the deepfake, raised significant concerns about the integrity of democratic information channels.

Figure 2

Deepfakes Induced Concerns Related to U.S. Democracy



Results are reported per each hypothesis below.

Hypothesis 1: Perceptions of Policy Effectiveness

To evaluate whether perceptions of AI regulation impact societal trust, participants were asked to respond to three policy-focused questions using a 5-point Likert scale (1 = Not at all / No trust, 5 = Extremely effective / Completely trust)[5].

Q11 asked participants how effective they believed content labeling laws would be in reducing the spread of AI-generated misinformation. The average score was 3.28, which was not statistically different from the neutral midpoint (t(n) = 1.81, p = .078). This suggests public ambivalence about the effectiveness of labeling laws alone as an effective policy.

Q12 asked participants whether they believed stronger regulations were needed to address technologies like deepfakes. This question received a mean score of 4.41, a highly

[5] See the Appendix for more on the questions and respective answers.

significant result (t(n) = 18.66, p < .001), indicating strong support for more robust regulatory measures.

Q13 asked about trust in the government to enact effective AI legislation. Participants gave this question an average score of 2.88, which was not significantly different from the midpoint (p = .635), suggesting neutral to slightly low trust in government competence to regulate AI responsibly.

As shown in Figure 3, while participants overwhelmingly support the need for stronger regulation, their trust in government institutions to implement such measures remains ambivalent. Labeling laws were also only met with cautious optimism.

Figure 3

Mean Scores for AI Policy Perception (Q11-Q13)

Note. Q12 (support for stronger regulation) received high ratings, while responses to Q11 (labeling laws) and Q13 (trust in government action) were closer to neutral.

Hypothesis 2: Deepfake Exposure and Trust Erosion

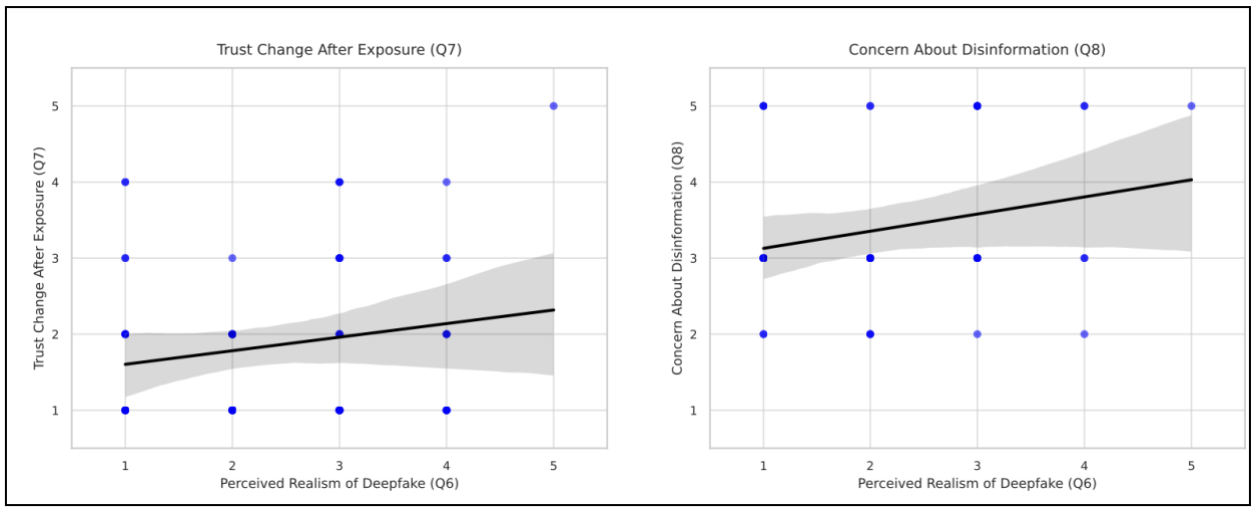
To assess the impact of deepfake exposure, this study assessed whether participants' perceptions of deepfake realism (Q6) were associated with changes in trust and concern about democratic integrity. Pearson correlation analyses were conducted between Q6 and four post-exposure items: trust erosion (Q7), concern about disinformation (Q8), privacy concern (Q9), and concern about democratic breakdown (Q10).

A weak but statistically significant correlation was observed between deepfake realism and trust erosion ($r = .25$, $p = .04$), suggesting that participants who perceived the deepfake as more realistic were more likely to report a decrease in trust in political figures (Q7). Additionally, marginally significant positive correlations emerged between deepfake realism and concern about disinformation ($r = .21$, $p = .08$; Q8) and concern about privacy ($r = .21$, $p = .08$; Q9); this indicates a potential trend of elevated concern among those more worried by the realism of AI-manipulated media.

As shown in Figure 4, the trendlines illustrate a positive association across trust erosion and disinformation outcome variables, with participants who rated the deepfake as more realistic also tending to report higher levels of institutional mistrust and concern. Each point on the scatterplots represents one or more participants; because responses were limited to a 5-point Likert scale, overlapping data points are common.

Figure 4

Deepfake Correlation on Trust Erosion and Disinformation Concerns

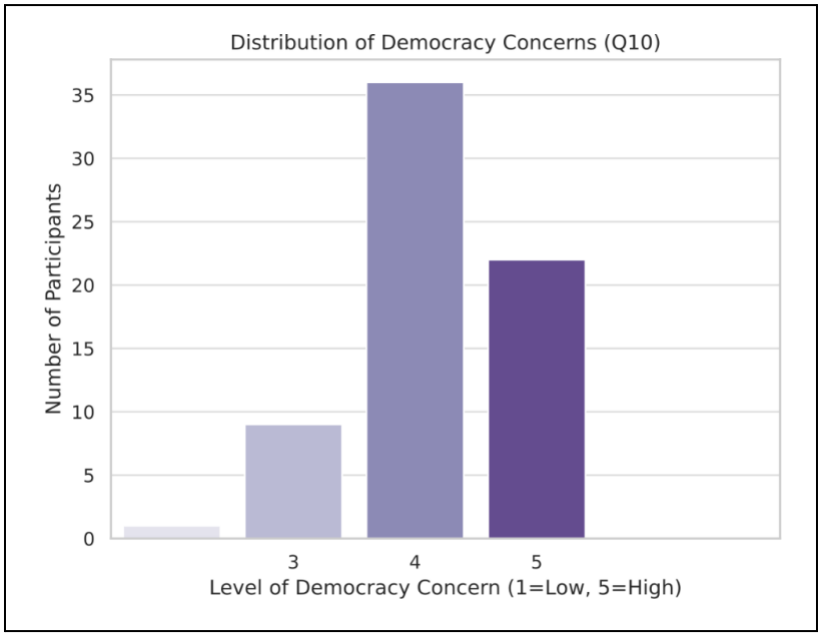


Note. Scatterplots displaying the relationship between perceived realism of the deepfake video (Q6) and post-exposure measures of trust erosion (Q7) and concern about disinformation (Q8).

The strongest relationship was observed between realism and trust erosion (r = .25, p = .04), with other indicators showing marginal or non-significant associations. Although no statistically significant relationship was found between deepfake realism and perceived threat to democracy (r = .19, p = .12; Q10), this result may reflect a ceiling effect more than a lack of concern. Over 85% of participants already agreed or strongly agreed that deepfakes pose a serious threat to U.S. democratic integrity. With such a high level of baseline concern, there was little room left for perceived realism to make a measurable difference. In this sense, the data still suggests that deepfake media poses a profound threat to public trust—as seen in Figure 5—, even if the realism of the specific clip did not significantly enhance such perceptions further.

Figure 5

Post Deepfake Exposure Democratic Integrity Concerns



Discussion

This study examined how exposure to an AI-generated deepfake influences trust in U.S. governmental institutions and public perceptions of AI regulatory principles. The findings offer a nuanced picture of deepfake realism and perceived concerns. Participants strongly supported the need for robust regulation to address AI risks, yet expressed mixed confidence in transparency-based solutions like content labeling. Although perceived realism of the deepfake video was only moderately associated with decreases in trust, concern about the societal risks of manipulated media was widespread. Even when participants rated the deepfake as relatively unrealistic, the overwhelming majority still expressed significant concern about threats to privacy, information authenticity, and democratic integrity. Together, these results suggest that public anxiety about deepfakes may operate independently of technical realism, pointing to a

deeper erosion of societal trust based on uncertain epistemic agency derived in the age of AI-generated content.

Public Support for Robust AI Regulation

Participants' perceptions of AI regulation closely align with the goals of California's proposed legislation. In particular, strong support for robust regulation (Q12) reflects public support of the foundational principles behind SB 1047, which aimed to introduce comprehensive oversight mechanisms for advanced AI development. Meanwhile, moderate skepticism about the effectiveness of content labeling alone (Q11) mirrors critiques of SB 942, which proposed mandatory disclosure of AI-generated content. Importantly, respondents' strong concern about threats to privacy (Q9) and authenticity (Q10) further validates the need for regulatory frameworks that not only address transparency, but also deeper questions of information integrity and epistemic agency. However, the neutral to slightly low trust in government capacity to enact such measures (Q13) suggests that even well-designed policies may struggle to restore public confidence unless accompanied by meaningful institutional reforms and increased credibility.

Deepfake Realism and the Erosion of Trust

Building on these regulatory findings, the experimental results related to deepfake exposure further highlight the complexity of the relationship between manipulated media and societal trust. Although perceived realism of the deepfake was not high, participants who rated the video as more realistic were significantly more likely to report decreased trust in public figures (Q7). The positive correlation between deepfake realism and trust erosion, albeit moderate ($r = .25$, $p = .04$), suggests that even modestly convincing deepfakes have the potential to weaken trust in democratic institutions. In addition, while correlations between realism and concerns about disinformation (Q8) and privacy infringement (Q9) were marginally significant,

they nonetheless point to an emerging trend: the more realistic manipulated media appears, the more it seems to amplify public anxiety about informational authenticity and epistemic agency.

Broader Anxiety About Deepfakes and Information Authenticity

However, the more striking finding lies not in the correlation analyses, but in the broader levels of concern participants expressed across the board. Even when deepfake realism was perceived as low, concern about disinformation (86% somewhat to extremely concerned), privacy infringement (84%), and threats to democratic integrity (85%) remained overwhelmingly high. This pattern suggests that the presence of deepfakes in our media ecosystem (even when relatively unrealistic) may be sufficient to destabilize public confidence in the reliability of information, if it has not already. In this sense, the societal risk posed by deepfakes is not merely a question of technological sophistication, but one of perceived possibility. It seems that once the public realizes that images, videos, and audio can be convincingly manipulated with little to no detection, the epistemic foundation of information authenticity is threatened.

Limitations

At the same time, this study has several limitations that warrant consideration. First, the sample was drawn from mostly college students at the University of California, Irvine due to funding restrictions and may be unrepresentative of the national population, which would limit the application of findings. Second, participants were exposed to a single deepfake of a well-known political figure, whose script included uncharacteristic use of profanity. Finally, ceiling effects related to the measures of concern may have dampened the ability to detect stronger correlations between deepfake realism and trust-related outcomes.

In addition to these methodological considerations, it is important to note a broader scope condition: the sample skewed young and likely reflected higher levels of digital literacy than the

general population. Most participants were undergraduate students, and their ability to correctly identify the deepfake may not be representative of older or less digitally fluent age groups. Future studies should explore how different generational cohorts respond to deepfakes, as lower detection ability among older users may compound the risks to institutional trust and further complicate regulatory efforts.

Implications

Future research could build on these results by examining how the frequency of exposure to deepfakes influences trust over time, whether exposure to different types of figures (e.g., less familiar or more polarizing) affects responses, and how interventions such as pre-bunking[6], content labels, or media literacy training might mediate deepfakes' impact on trust and epistemic agency. In addition, studies that explore partisan or demographic differences in susceptibility to deepfake-related trust erosion could enhance our understanding of how these threats manifest across a diverse democratic public.

Policy Recommendations

In addition to future research directions, the findings also point to several concrete policy recommendations aimed at restoring trust and curbing the harms of deepfake media. Based on this study's findings, several policy recommendations emerge to address the challenges posed by deepfake technology. First, while content labeling laws such as California's SB 942 received moderate support, the data suggests that transparency alone may not be sufficient to preserve public trust. Future legislation should move beyond labeling to include stronger accountability measures for AI developers, as envisioned in SB 1047. Such frameworks could require safety audits, disclosure of training data sources, and ethical oversight for high-risk AI systems.

[6] Pre-bunking is a psychological inoculation strategy that exposes people to weakened examples of misinformation tactics in advance, helping them build resistance against future manipulation (Roozenbeek & van der Linden, 2019).

Second, given participants' overwhelming concern about disinformation and privacy threats—regardless of how realistic the deepfake appeared—lawmakers should consider proactive public education campaigns. These might include media literacy programs, deepfake detection training, or collaborative efforts between government and tech companies to build public resilience.

Finally, the finding that most participants supported strong regulation but had low trust in government to enact it highlights the need for greater transparency in the regulatory process itself. Efforts to engage the public, clarify policy intent, and demonstrate responsiveness may be crucial to restoring both informational and institutional trust in an AI-saturated media environment.

Conclusion

Taken together, the findings from this study point to an urgent reality: while deepfakes may not always succeed in deceiving individuals outright, their existence contributes to a broader erosion of informational certainty. Public support for robust regulatory action is strong, but skepticism about transparency measures and government effectiveness reveals a gap that legislation alone cannot bridge. In a world where technology can blur lines of truth, safeguarding democratic trust will require not only well-crafted policy, but also a renewed cultural investment in critical thinking, civic education, and collective responsibility for informational integrity.

REFERENCES

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency* (pp. 149–159). https://doi.org/10.1145/3287560.3287583

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2018). *OpenAI gym*. arXiv preprint arXiv:1606.01540. https://arxiv.org/abs/1606.01540

BuzzFeed. (2018). *You won't believe what Obama says in this video!* YouTube. https://www.youtube.com/watch?v=cQ54GDm1eL0

California Consumer Privacy Act of 2018, Cal. Civ. Code §§ 1798.100–1798.199 (West 2020). https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5

California Governor's Office. (2024, September). *Governor Newsom's veto statement on SB 1047 and signing statement on SB 942*. Office of Governor Gavin Newsom.

Chesney, R., & Citron, D. K. (2019). Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Affairs, 98*(1), 147–155.

Coeckelbergh, M. (2024). Deepfakes and the politics of knowledge: Epistemic risks and democratic governance. *AI & Society, 39*(1), 3–12. https://doi.org/10.1007/s00146-022-01419-1

Dan, V., Ceron, A., & Osnabrügge, A. (2021). The impact of deepfake videos on political attitudes: An experimental study. *New Media & Society, 23*(10), 2933–2955. https://doi.org/10.1177/1461444821992285

Devos, T., Spini, D., & Schwartz, S. H. (2002). Conflicts among human values and trust in institutions. *British Journal of Social Psychology, 41*(4), 481–494. https://doi.org/10.1348/014466602321149849

European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)*. *Official Journal of the European Union*, L119, 1–88. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Gilbert, M. (1987). *On social facts.* Princeton University Press.

Hawley, K. (2014). Trust, distrust and commitment. *Nous, 48*(1), 1–20. https://doi.org/10.1111/j.1468-0068.2012.00889.x

Heimer, C. A. (2001). Solving the problem of trust. In K. S. Cook (Ed.), *Trust in society* (pp. 40–88). Russell Sage Foundation.

Hetherington, M. J. (2005). *Why trust matters: Declining political trust and the demise of American liberalism.* Princeton University Press.

Meta. (n.d.). *Ad Standards: Social issue, electoral, and political advertising.* Meta Transparency Center. Retrieved April 26, 2025, from https://transparency.meta.com/policies/ad-standards/siep-advertising/siep

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications, 5*(1), 1–10. https://doi.org/10.1057/s41599-019-0279-9

U.S. Congress. (2024). *S.3875 – AI Transparency in Political Ads Act, 118th Congress (2024).* Congress.gov. Retrieved April 26, 2025, from https://www.congress.gov/bill/118th-congress/senate-bill/3875/text

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), 1–13. https://doi.org/10.1177/2056305120903408

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review, 41*, 105567. https://doi.org/10.1016/j.clsr.2021.105567

Yanamala, N., Arora, D., & Patel, A. (2023). Regulating AI: Challenges and opportunities for transparency and accountability. *Journal of Law and Technology Policy, 2023*(1), 45–72.

Appendix A: Full Survey Instrument and Question Wording

Survey Questions

Section I: Pre-Exposure Baseline Questions (All Participants)	
Questions	**Likert Scale Answer Choices**
Q1. How much do you trust U.S. political institutions (e.g., government, political leaders) to provide truthful information?	1 = No trust, 5 = Completely trust
Q2. How familiar are you with the concept of deepfakes?	1 = Not familiar, 5 = Extremely familiar
Q3. How concerned are you about the spread of disinformation in U.S. media?	1 = Not concerned, 5 = Extremely concerned
Q4. How effective do you believe current laws are in protecting citizens from the misuse of AI technologies like deepfakes?	1 = Not at all effective, 5 = Extremely effective
Q4b. If you answered "Not at all" or "Slightly effective" please provide brief reasoning; if not, simply put "N/A"	[text input]
Section II: Video Evaluation	
Q5. Which of the two videos seemed less authentic to you?	Options: Video 1 / Video 2
Q5b. Which of the two videos do you believe was a deepfake?	Options: Video 1 / Video 2 / I'm not sure
Q6. How realistic did the second video (Clip 2) seem to you?	1 = Not realistic at all, 5 = Extremely realistic
Section III: Post-Exposure Questions (All Participants)	
Q7 Group Hotdog. How much do you trust U.S. political institutions (e.g., government, political leaders) to provide truthful information?	1 = No change, 5 = Complete loss of trust
Q7 Group Hamburger. After watching these videos, how has your trust in statements from public figures changed?	1 = No change, 5 = Complete loss of trust

Q8. How concerned are you about manipulated media (e.g., deepfakes) in spreading disinformation?	1 = Not concerned, 5 = Extremely concerned
Q9. Do videos like these increase your concerns about privacy infringement?	1 = Not at all, 5 = Extremely concerned
Q10. Do videos like these raise concerns about threats to authenticity and privacy in U.S. democracy?	1 = Strongly disagree, 5 = Strongly agree
Section IV: AI Policy Perception Questions (All Participants)	
Q11. How effective would laws requiring companies to label AI-generated content be in reducing the spread of misinformation?	1 = Not at all effective, 5 = Extremely effective
Q12. Do you believe there is a need for stronger regulations to address AI technologies like deepfakes?	1 = Strongly disagree, 5 = Strongly agree
Q13. How much do you trust the government to enact laws that effectively address the risks of AI technologies?	1 = No trust at all, 5 = Completely trust
Q14. Do you think deepfakes pose a significant threat to the integrity of U.S. elections?	1 = Strongly disagree, 5 = Strongly agree
Q15. After this experiment, how confident are you in the authenticity of digital media (videos, audio, images)?	1 = Not confident at all, 5 = Completely confident
Section VI: Demographic Questions	
Q15. Age (short answer)	[text input]
Q16. Do you primarily reside in California?	(Yes/Other: [text input])
Q17. How often do you follow political news?	1 = Never, 5 = Always
Q18b. What is your primary source for such news?	Options: TikTok / Instagram / TV / Radio / Other: [text input]