AI and Deepfakes: An Examination of Educational Interventions in Combating Deepfake-Based Disinformation

Marc J. Dupuis
Computing & Software Systems
University of Washington
Bothell, USA
marcjd@uw.edu

Hyunyoung Sung
Computing & Software Systems
University of Washington
Bothell, USA
hyunsung@uw.edu

Sophia Long
Computing & Software Systems
University of Washington
Bothell, USA
sophiamailong@gmail.com

Abstract—**Deepfake technology has rapidly emerged as both an innovation in synthetic media and a significant societal threat, undermining trust in healthcare, climate science, politics, and beyond. While technical detection tools are advancing, less is known about the effectiveness of educational interventions designed to improve human judgment in distinguishing between authentic and manipulated content. This study investigates whether targeted training can enhance individuals' ability to detect deepfake images. A large-scale survey was conducted with 335 participants recruited via Amazon's Mechanical Turk and randomly assigned to treatment and control groups. All participants first completed a pre-test assessing their ability to identify real and deepfake images across political, climate, and health domains. The treatment group then viewed an educational video on deepfake detection, while the control group viewed neutral content. Both groups subsequently completed a post-test with an expanded image set. Results show that participants who received the educational intervention significantly outperformed the control group in detecting deepfake images, supporting our first hypothesis (H1). However, the same group performed worse in identifying authentic images, leading to the rejection of our second hypothesis (H2). These findings reveal an important trade-off: while education can improve skepticism toward synthetic media, it may also foster over-skepticism and reduce trust in genuine content. This study contributes empirical evidence to the debate on human-centered countermeasures against disinformation, highlighting both the promise and the pitfalls of educational approaches. We argue that effective solutions will require integrated strategies combining education, feedback mechanisms, and platform-level detection systems.**

Index Terms—**synthetic media, deepfake images, educational interventions, disinformation, human-centered security, media trust**

I. Introduction

Recent advancements in artificial intelligence (AI) and machine learning (ML) have paved the way for the development of deepfake technology, an advanced tool capable of producing highly realistic synthetic media. The early phase of AI-generated image manipulation marked the beginnings of deepfakes, which were further advanced through machine learning models. A major breakthrough came with the introduction of Generative Adversarial Networks (GANs) in 2014, a neural network architecture designed to generate new data from existing datasets. NVIDIA's StyleGAN exemplifies the application of GANs by producing realistic human faces for use in video game avatars and marketing campaigns [1]. By 2017, deepfake technology had become widely accessible to the public, raising concerns about social manipulation, identity theft, and disinformation campaigns, and contributing to an explosion of AI-generated content.

As consumption of synthetic media has increased, so too have concerns about hyper-realistic misinformation, particularly in the fields of healthcare, climate science, and politics. In climate science, AI has supported weather prediction, environmental monitoring, and disaster forecasting by analyzing large-scale datasets from satellites and sensors [2]. It has also contributed to renewable energy optimization and conservation efforts. Yet, the rise of "deepfake geography" has enabled the creation of falsified satellite images, spreading false narratives about climate change, manipulating public opinion, and undermining scientific evidence.

Similar dynamics are evident in healthcare. AI tools have been applied to analyze vast quantities of medical data, detect patient vulnerabilities, create therapeutic virtual reality videos for Alzheimer's patients [3], and improve diagnostic imaging in oncology and pathology. However, deepfakes have also been misused to generate misleading medical claims, fabricated research images, and manipulated patient testimonials. Such fraudulent practices can mislead healthcare professionals, distort research, and foster public misunderstanding, with downstream effects on health policy.

In the political sphere, deepfakes have eroded public trust by enabling the fabrication of audio and video recordings of political figures. Unlike traditional video manipulation tools, deepfake software is often free and unlicensed, allowing almost anyone to create persuasive fabrications [4]. One widely cited example is the 2018 deepfake of Barack Obama, which illustrated the potential of synthetic media to mislead the public and destabilize political discourse. The growing accessibility of such tools underscores the urgent need for robust detection methods, public education, and protective measures against disinformation [4].

Taken together, these developments demonstrate that while advances in AI have enabled constructive applications, the misuse of deepfakes has fueled misinformation, public skepticism, and erosion of confidence in authentic media [2], [5]. A growing body of research highlights the difficulty individuals face in distinguishing real from fabricated content, underscoring the need for effective interventions that improve detection accuracy.

To address this challenge, the present study examines the efficacy of educational interventions in improving individuals' ability to distinguish between real and synthetic images. Specifically, we investigate whether a targeted training intervention enhances participants' ability to discern both deepfake and authentic images. This study evaluates two hypotheses:

- **H1:** Participants in the treatment group that undergo the educational intervention will more accurately identify deepfake images than those in the control group.
- **H2:** Participants in the treatment group that undergo the educational intervention will more accurately identify real images than those in the control group.

By testing these hypotheses, this study contributes empirical evidence on the effectiveness of educational interventions as a countermeasure to deepfake disinformation. The following section reviews related work on deepfakes in healthcare, climate science, and politics, situating the current study within the broader literature.

II. Background

In an era where digital media dominate public discourse, deepfake technology has ushered in a new form of synthetic deception. What began as an innovative application of AI for entertainment has rapidly evolved into a cutting-edge manipulation tool that blurs the boundaries between reality and fabrication. Deepfakes therefore represent a double-edged sword—an advanced tool with the potential to revolutionize digital content, yet equally capable of undermining trust in climate science, healthcare, and politics [5].

Disinformation produced through AI has raised widespread concerns among healthcare professionals, policymakers, and technologists. While AI has delivered benefits in domains such as policymaking, environmental conservation, and medical diagnostics, deepfakes complicate the authentication of climate data [2], erode public trust in healthcare, and sow confusion in democratic processes. Given the pace of AI development and the accessibility of deepfake tools, it is essential to examine the causes, effects, and countermeasures of misinformation. This section reviews prior work on the role of deepfakes in healthcare, climate science, and politics, highlighting both constructive applications and risks.

A. Deepfake Misinformation in Healthcare

AI-driven healthcare innovations have enhanced diagnostics, treatment strategies, and patient engagement. Deepfake technology builds on these developments, with applications ranging from medical training to patient education. For example, realistic patient personas have been used to simplify complex medical concepts, improve engagement, and support treatment compliance [5]. Similarly, AI-generated avatars have assisted patients with panic and anxiety disorders by enabling them to rehearse social interactions in safe, simulated environments. Popular media such as the film *CTRL* (2024) dramatize the potential harms of such technologies, underscoring the risks of identity manipulation and psychological distress [6]. Telemedicine has also adopted deepfake tools: Synthesia has developed avatars that anonymize patient identities while maintaining expressions, providing a layer of privacy in remote consultations [7]. While these applications promise greater access and compliance with privacy regulations such as HIPAA, they simultaneously raise concerns regarding data security, informed consent, and the authenticity of medical interactions [5].

Beyond clinical care, medical education has also benefited from deepfakes. Institutions such as Johns Hopkins University employ AI-driven avatars for nurse training, describing them as "the next-best thing" to real patients [8], [9]. The Mayo Clinic has used synthetic datasets to improve radiologists' ability to detect tumors, while simultaneously protecting sensitive patient information. These cases demonstrate the constructive potential of deepfake applications in healthcare.

Despite these benefits, the risks are significant. Vaccine hesitancy has been fueled by fabricated videos spreading misinformation about vaccine safety [5], [10]. During the COVID-19 pandemic, Russia disseminated deepfake content alleging severe side effects and infertility linked to U.S. vaccines such as Pfizer, reinforcing public distrust [11], [12]. Fabricated patient testimonials have also been employed in fraudulent advertising, as seen in manipulated videos of well-known doctors such as Michael Mosley and Rangan Chatterjee promoting false cures for diabetes [13], [14]. Even the integrity of peer-reviewed science has been compromised. The Surgisphere scandal, which involved manipulated data published in *The Lancet* and *The New England Journal of Medicine*, misled global COVID-19 treatment protocols before being retracted [15], [16]. These incidents underscore the dangers deepfakes pose to public health and research credibility.

B. Deepfake and Climate Disinformation

AI and machine learning have significantly advanced climate modeling, disaster forecasting, and renewable energy optimization. At the same time, deepfake technology has created new avenues for climate-related misinformation. On the constructive side, activists in high-risk regions have used deepfake facial modifications to protect their identities while exposing evidence of deforestation, pollution, and other environmental threats. AI-generated climate models, such as the European Union's Destination Earth project, aim to build digital twins of the planet that replicate natural and human systems, thereby improving disaster preparedness and policymaking [17]. Similarly, NVIDIA's FourCastNet applies neural operators to predict extreme weather [18], and Google's DeepMind has optimized renewable energy forecasts to enhance grid efficiency and reduce fossil fuel dependence [19]. Counterbalancing these positive uses, deepfakes have also undermined climate communication. Fabricated reports and manipulated videos of scientists have been used to distort public understanding of climate change. In 2023, AI-generated images were circulated to falsely suggest that offshore wind turbines were responsible for whale deaths, fueling opposition to renewable energy initiatives [20]. Deepfakes have also been deployed to generate false weather warnings; in 2021, a viral TikTok video claimed a Category 5 hurricane was about to strike Miami, creating widespread panic and even short-term market impacts before being debunked [21]. Manipulated satellite datasets further complicate public discourse, such as a 2022 viral article that falsely claimed Arctic ice was at a 30-year high [22]. Such examples illustrate how fabricated content can undermine environmental science, policymaking, and public trust.

C. Deepfake Influence in Political Media

Deepfakes have profound implications for governance, public opinion, and democratic integrity. They can be used ethically, such as in political awareness campaigns designed to educate citizens about misinformation. For example, companies have produced labeled deepfakes of political leaders to demonstrate how easily speeches can be altered, thereby raising media literacy [23]. French President Emmanuel Macron even appeared in humorous deepfake videos in 2025, where his likeness was placed in popular film scenes, to promote an AI summit in Paris and highlight the technology's potential [24]. Deepfakes have also been used to increase accessibility, as when India's BJP party translated candidate Manoj Tiwari's speech into Haryanvi during the 2020 elections, broadening his reach to local voters [25]. In authoritarian contexts, activists and journalists have considered using deepfake facial modifications to avoid surveillance while still sharing critical messages, such as during Hong Kong's pro-democracy protests [26].

However, the risks in the political sphere are especially acute. Deepfake videos of political leaders making false statements can mislead the public, damage reputations, and influence elections. In 2024, a deepfake robocall of U.S. President Joe Biden urged voters in New Hampshire not to participate in primaries, constituting an illegal voter suppression tactic [27]. Fabricated endorsements and manipulated recordings have also destabilized political systems, as seen in Slovakia in 2023, when deepfake audio implicated opposition leaders in election-rigging schemes [28]. More broadly, deepfakes contribute to what scholars call the "liar's dividend" [29], where authentic content is dismissed as fake. This dynamic was evident in Australia's 2024 federal election, when a politician claimed that a damaging video was a deepfake, prompting forensic verification and fueling public mistrust [10]. These examples illustrate how deepfakes exacerbate uncertainty in political communication and complicate accountability in democratic systems.

D. Conclusion

The literature reveals that deepfake technology is both a transformative innovation and a profound societal risk. Across healthcare, climate science, and politics, deepfakes demonstrate clear potential for constructive uses—enhancing education, protecting identities, and expanding access to information. At the same time, they enable misinformation campaigns, fabricated research, and manipulated media that erode public trust, compromise democratic processes, and endanger public health.

A consistent theme across studies is the dual-use nature of deepfakes: ethical applications are often shadowed by parallel opportunities for misuse. This tension underscores the importance of developing technical safeguards, regulatory frameworks, and public education to mitigate harms while supporting legitimate innovation. Importantly, prior work demonstrates that individuals themselves cannot reliably detect deepfakes, and that existing interventions such as awareness-raising or financial incentives do little to improve detection accuracy [30]. Instead, participants exhibit both authenticity bias and overconfidence in their judgments, further amplifying their vulnerability to manipulation.

Taken together, the existing body of work highlights the urgent need for new, human-centered approaches to address the growing challenges posed by deepfake technology. These insights form the foundation for the present study, which tests the effectiveness of an educational intervention designed to improve detection of both deepfake and authentic images.

III. Methods

In this study, we address the underlying research question by conducting a large-scale survey. Institutional Review Board (IRB) approval was sought and obtained prior to participant recruitment. Additionally, informed consent was received prior to the engagement of any participants in the study. Amazon's Mechanical Turk (MTurk) was employed to recruit participants. While MTurk has had significant issues with quality over the years, if several measures are employed then said issues may largely be mitigated [31]–[33]. The survey itself was hosted using the Qualtrics survey platform.

There were 404 participants that began the survey with 28 excluded due to being affiliated with the university conducting the study. Another 71 participants failed one or more quality control questions embedded in various parts of the survey with 335 participants successfully completing it. The overall quality control failure rate of 18.9%. Participants were compensated $3 for their time and effort with 92.5% of them indicating that the compensation received was comparable or easier for the money when compared to similar projects on MTurk.

The participants were randomly assigned to either the control group or treatment group by the Qualtrics survey platform. Each group participated in a pre-tets, which consisted of them assessing six images to identify whether they were real or

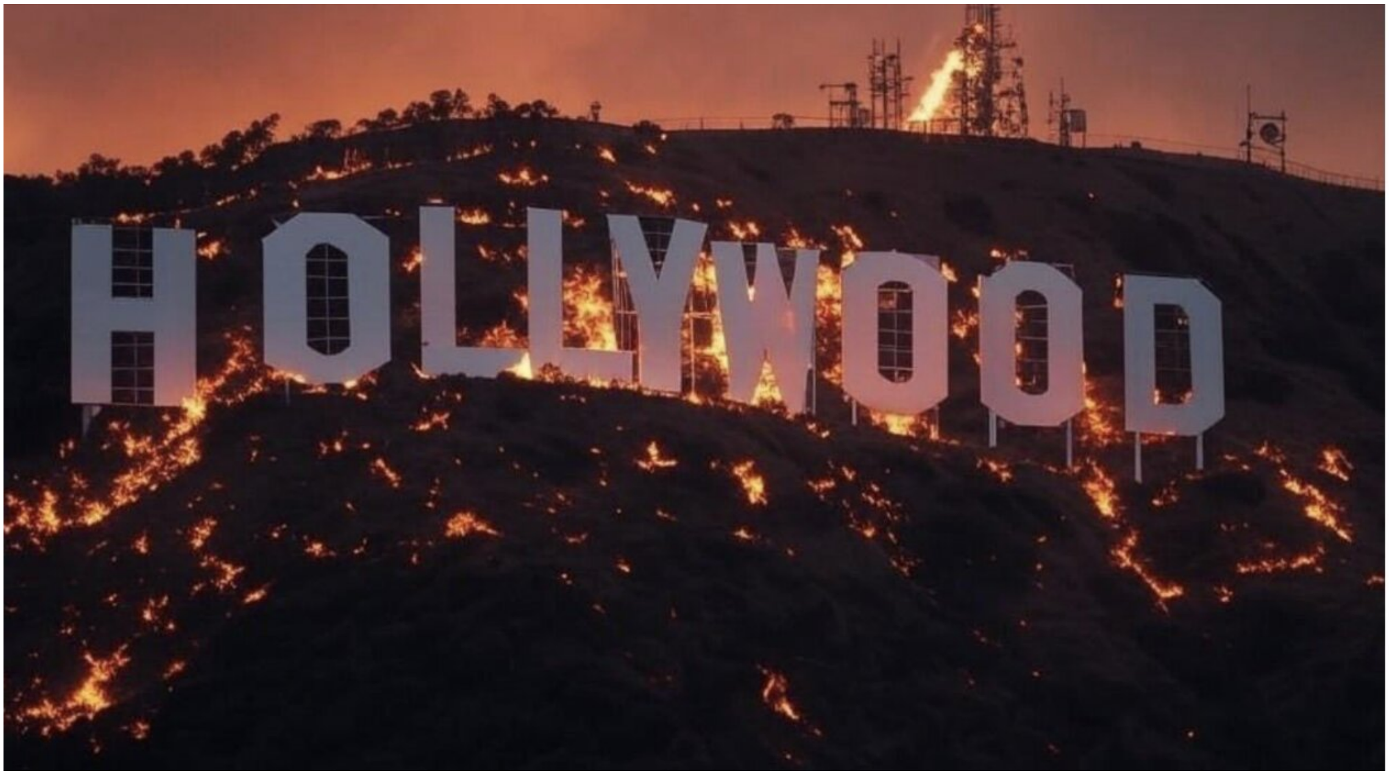

Fig. 1. Deepfake related to climate

deepfakes with an equal number of each presented to the participants in a random order. The images came from three different categories–political, climate, and health–with both a deepfake and real image from each presented to the participants.

Next, the control group watched an innocuous video of scenery with neutral music in the background, while the treatment group was exposed to a video that sought to educate them on how to identify a real image from a deepfake image. Upon completion of the videos, participants in both groups were then assessed with an additional series of images to evaluate. This time, there were 18 images total for them to evaluate with 6 images from each of the three categories evenly split between real and deepfakes. It is important to note that a mixture of easy to identify and more challenging deepfake versus real images were used. And while they were not tasked with comparing images in a similar context, the following images illustrate what at times can be a stark contrast between the two (see Fig. 2 and Fig. 3).

IV. Results and Analysis

A. Demographics

Of the 335 participants who successfully completed the survey, 55.8% identified as male, 43. 6% as female, 0. 3% as non-binary / third gender, and 0. 3% did not wish to answer the question. A large majority of the participants were White (79.1%), followed by Asian/Pacific Island (10.1%), Black/African-American (6.3%), Hispanic (3.3%), Other/Multi-Racial (0.9%), and Native American/Alaskan Native (0.3%). Almost half (49.3%) had a Bachelor's degree or higher, while only 13.7% had no schooling beyond graduating high school or the equivalent.

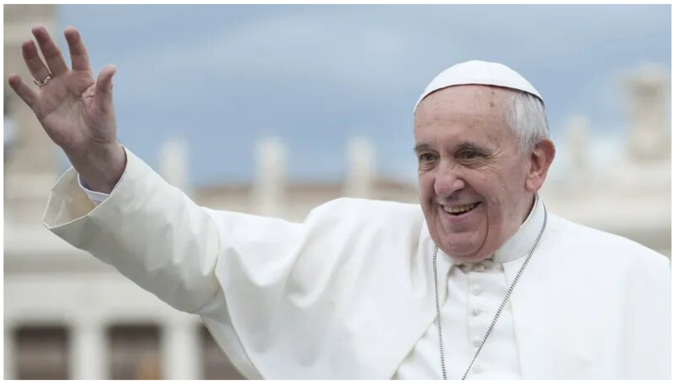

Fig. 2. Real image of a political nature

B. General Question Responses

Participants were asked a series of questions related to deepfakes so that we could better understand their baseline level of experiences and perceptions related to deepfakes.

- When using any form of social media, how often have you encountered media that you believed to be deep fake or ai-generated?

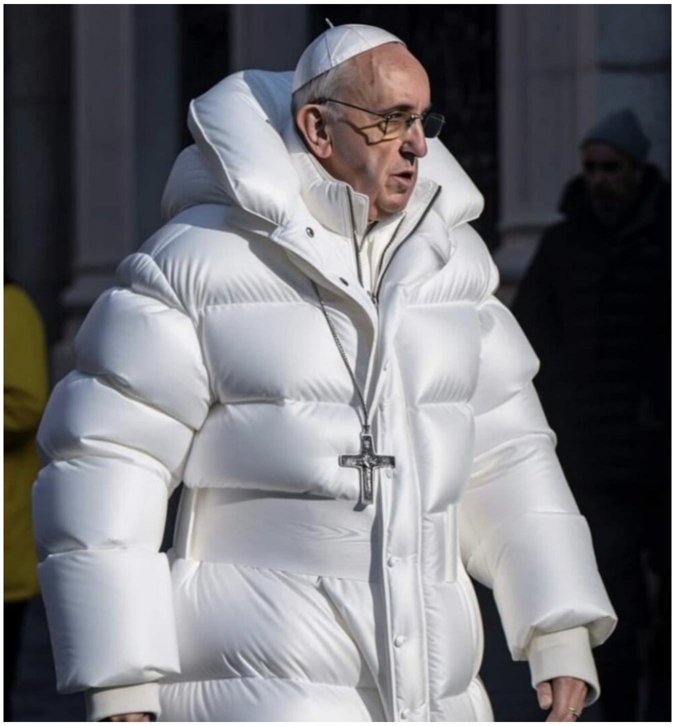

Fig. 3. Deepfake image of a political nature

Most participants (69.6%) indicated that they had encountered deepfake or AI-generated media a moderate amount, followed by those that often encountered it (20.3%), or believe they have never encountered it (10.1%). Again, this is their perception of what they had encountered–not what they had actually encountered. This relates quite closely to our next question.

- How confident are you in your ability to detect any form of synthetic media?

A large percentage of participants (57.9%) felt quite confident in their ability to detect synthetic media, while 28.1% made a selection that suggested they were not confident and only 14.0% noted they were highly confident.

- How much do deepfakes affect your trust in news and online media?

Approximately one third (33.7%) of participants indicated that deepfakes moderately impacted their trust in news and online media. Almost as many (30.4%) reported a considerable impact, while 14.3% suggested a significant loss of trust. Only 21.5% indicated either no impact or slight impact.

- Have you ever questioned the authenticity of a government or political figure's statement because of deepfake concerns?

The questioning of statements made by governmental or political figures is a growing concern and this was reflected by our participants as well. A significant number (61.8%) of them noted that they have questioned the authenticity of statements due to deepfake concerns. While many (28.4%) indicated that they have not or were not sure (9.6%), the additional loss of trust in this area due to changes in technology is a concern.

- Do you trust videos or images sent by friends and family, or do you feel the need to verify them first?

Despite their reduction in trust in statements made by political figures due to deepfakes, only 26.6% of participants indicated that they always verify videos or images sent by friends, while 59.7% do so sometimes. A small subset (13.7%) appear to have complete faith in what they receive from their friends.

- What concerns you the most about deepfakes?

Participants noted a variety of concerns related to deepfakes with misinformation being most prominent (66.0%), followed by political manipulation (20.9%), privacy violations (4.8%), identity theft (4.2%), falsified health related information (2.7%), among other concerns (1.5%).

- Do you believe social media platforms should be responsible for detecting and labeling deepfake content?

Finally, there was a strong consensus (76.7%) that social media platforms should be responsible for detecting and labeling deepfake content. Only 10.7% of participants did not believe they should be responsible, while 12.5% were not sure.

C. *Educational Intervention*

Prior to assessing the efficacy of the educational intervention, we evaluated the results of the pre-test to determine if the scores of the control and treatment groups were statistically different by performing an independent samples t-test. The results indicated that the performance of both the control and treatment groups were not different in any meaningful way for our purposes here. Therefore, we could then proceed with further analyses to assess the educational intervention received and its efficacy compared to the control group.

We began by calculating the score for the total number of correct responses obtained in the assessment after participants watched the video, whether in the control or treatment group. An independent samples t-test was performed to evaluate whether there was a difference between the number of correct answers of those in the treatment group and those in the control group. The results indicated that there was no significant difference between the number of correct answers of the treatment group (M = 11.35, SD = 2.20) and the control group (M = 11.45, SD = 2.12).

However, when we evaluated the ability of participants to accurately detect deepfakes and real images independently of one another, interesting differences were observed. An independent samples t-test was performed to evaluate whether there was a difference between the number of correct answers related to detecting deepfake images of those in the treatment group and those in the control group. The results indicated that the treatment group (M = 6.02, SD = 1.78) had a significantly higher score than the control group (M = 5.19, SD = 1.85), t(333) = -4.213, $p < .001$. Therefore, H1 is supported.

Interestingly, the opposite was true as it related to their ability to detect real images. An independent samples t-test was

TABLE I
INDEPENDENT SAMPLES T-TEST RESULTS WITH CONFIDENCE INTERVALS AND EFFECT SIZES

Comparison	**Group**	**M**	**SD**	**95% CI**	**Cohen's** d	**p**
Total correct	Treatment	11.35	2.20	[10.97, 11.73]	-0.05	.74
	Control	11.45	2.12	[11.08, 11.82]		
Deepfake detection	Treatment	6.02	1.78	[5.73, 6.31]	0.45	$< .001$
	Control	5.19	1.85	[4.89, 5.49]		
Real image detection	Treatment	5.33	1.96	[5.01, 5.65]	-0.49	$< .001$
	Control	6.26	1.76	[5.96, 6.56]		
Gender (real images)	Male	6.01	1.98	[5.68, 6.34]	0.26	.018
	Female	5.51	1.79	[5.22, 5.80]		

performed to evaluate whether there was a difference between the number of correct answers related to detecting real images of those in the treatment group and those in the control group. The results indicated that the treatment group (M = 5.33, SD = 1.96) had a significantly lower score than the control group (M = 6.26, SD = 1.76), t(333) = 4.581, p <.001. Therefore, H2 can be rejected.

Finally, we compared the results based on gender identification. An independent samples t-test was performed to evaluate whether there was a difference between the number of correct answers related to detecting real images of those that identified as male and those that identified as female regardless of treatment or control group status. The results indicated that those that identified as male (M = 6.01, SD = 1.98) had a significantly higher score than those that identified as female (M = 5.51, SD = 1.79), t(331) = 2.375, p = .018. The significance of this finding is unclear, especially given the lack of significant difference between gender identification as it relates to the ability to detect deepfakes.

V. Discussion

This study investigated the efficacy of an educational intervention designed to improve individuals' ability to discern deepfake images from authentic ones. Two hypotheses were tested: H1 proposed that participants in the treatment group would more accurately identify deepfake images than those in the control group, while H2 proposed that participants in the treatment group would more accurately identify real images than those in the control group. The results provide partial support for our expectations. Participants exposed to the educational intervention demonstrated significantly higher accuracy in detecting deepfake images, supporting H1. However, they simultaneously performed worse at detecting real images, leading to a rejection of H2 [34].

These findings are consistent with prior work suggesting that people are not naturally adept at identifying deepfakes, even when motivated by awareness campaigns or financial incentives [30]. The current results extend this line of research by showing that a short, targeted educational intervention can indeed improve deepfake detection, but at a cost: participants became more skeptical overall, resulting in reduced accuracy when evaluating authentic images. This trade-off illustrates the tension between the "liar's dividend" and the "seeing is believing" heuristic described in prior literature [35], [36]. Training appears to shift participants away from default credulity but may simultaneously increase over-skepticism, where even genuine images are doubted.

The gender-related finding adds another layer of complexity. While men in the sample demonstrated significantly higher accuracy than women when detecting real images, no gender differences were observed in deepfake detection. The reasons for this pattern are not immediately clear, and the result should be interpreted cautiously. It may reflect differences in confidence, prior exposure, or other unmeasured variables rather than inherent differences in detection ability [30]. Further research is necessary to explore whether gender consistently moderates deepfake detection performance or whether this finding is sample-specific.

Several limitations must also be acknowledged. First, the study relied on MTurk participants, which, despite quality control measures, may not be fully representative of the general population. Second, the intervention was limited to a short video, which may not provide sufficient depth or practice for long-term skill development. Third, while images from three domains (health, climate, and politics) were included, the generalizability of findings to other types of media, such as video or audio deepfakes, remains uncertain. Fourth, issues related to social desirability bias and common method bias cannot be fully ruled out [37]–[40]. Finally, the reduction in real image detection suggests a potential unintended consequence of training interventions that requires further exploration [41].

A. Practical Implications

The findings of this study carry several important implications for practitioners, educators, and policymakers seeking to combat deepfake-based disinformation. First, the positive effect of the intervention on deepfake detection suggests that targeted training can equip individuals with tools to critically evaluate synthetic media. However, the concurrent reduction in real image detection highlights the risk of fostering over-skepticism, which could undermine trust in authentic content. This outcome cautions against deploying educational campaigns in isolation or without careful calibration.

To be effective, interventions should strike a balance between promoting vigilance and preserving confidence in genuine media. One potential avenue is to pair educational training with feedback mechanisms that help participants learn to distinguish specific deepfake artifacts over time [42], [43]. Another is the integration of AI-assisted detection tools to complement human judgment, reducing reliance on subjective perception alone. For policymakers, the results suggest that public education initiatives must be complemented by platform-level detection and labeling systems to mitigate the broader societal risks posed by deepfakes.

B. Future Research

While this study offers new insights into the potential and limitations of educational interventions, several avenues for future research remain. First, subsequent studies should examine the long-term effectiveness of training. It is unclear whether the gains in deepfake detection observed here persist over time or whether they diminish without continued practice and reinforcement. Longitudinal designs and repeated interventions could help answer this question.

Second, future research should explore alternative training formats. The present intervention relied on a single short video, but more interactive approaches—such as gamified exercises, repeated exposure with feedback, or hands-on workshops—may better balance vigilance with trust in authentic content. Comparative studies could evaluate which types of interventions are most effective across different populations.

Third, further work is needed to assess generalizability. This study focused on still images, but deepfake technology spans multiple modalities, including video and audio. Understanding whether training effects extend across these modalities is critical for developing comprehensive countermeasures. Similarly, future work should consider diverse populations beyond Mechanical Turk samples to ensure findings apply more broadly.

Finally, the observed gender difference in detecting real images warrants further investigation. Although the result should be interpreted cautiously, exploring whether demographic or experiential factors shape detection ability could provide valuable insights for tailoring interventions. Together, these directions highlight the importance of continued interdisciplinary research into both the human and technological aspects of deepfake detection.

In summary, this study demonstrates that educational interventions can improve detection of deepfake images but may simultaneously impair the recognition of authentic ones. These results underscore the dual-use challenge of interventions, mirroring the dual-use nature of deepfakes themselves. Future work should seek to refine educational strategies in order to strike a balance between improving skepticism of manipulated content and maintaining trust in genuine media.

VI. Conclusion

Deepfake technology represents both a remarkable innovation in synthetic media and a pressing societal challenge. As this study has shown, individuals often struggle to accurately identify manipulated content, and educational interventions, while promising, carry complex trade-offs. Our findings indicate that targeted training can improve the detection of deepfake images, supporting H1. However, this improvement comes at the expense of accuracy in detecting authentic images, leading to a rejection of H2. This outcome underscores the delicate balance between fostering critical vigilance and preserving public trust in genuine media.

The results contribute to the growing body of literature on human-centered strategies to mitigate the risks of deepfake disinformation. Practically, they suggest that educational efforts alone may not be sufficient; interventions must be carefully designed, combined with feedback mechanisms, and complemented by platform-level detection and labeling systems. Scholarly, the findings highlight the need for continued interdisciplinary research into how humans perceive and respond to synthetic media, the durability of training effects, and the interplay between cognitive biases and emerging AI tools.

In sum, this study demonstrates both the potential and the limitations of educational interventions in combating deepfake-based disinformation. Addressing this challenge will require integrated solutions that combine education, technology, and policy. By refining these strategies, it may be possible to reduce the harms of deepfake media while supporting informed public discourse and strengthening trust in authentic digital communication.

References

[1] NVIDIA, "This person does not exist," 2024.

[2] P. M. G. Gnanaguru, "Artificial intelligence for climate sustainability: A comprehensive review of applications, challenges, and future prospects," in *Proceedings of the IEEE Conference on [Conference Name]*, pp. 1–10, IEEE, 2023.

[3] M. A. Taha, W. M. Khudhair, A. M. Khudhur, O. A. Mahmood, Y. I. Hammadi, R. S. A. Al-husseinawi, and A. Aziz, "Emerging threat of deep fake: How to identify and prevent it," in *Proceedings of the 6th International Conference on Future Networks & Distributed Systems*, (Tashkent TAS Uzbekistan), p. 645–651, ACM, Dec. 2022.

[4] W. A. Galston, "Is seeing still believing? the deepfake challenge to truth in politics," 2024. Accessed 16 Mar. 2025.

[5] J. Smith and J. Doe, "An innovative approach to iot security," *IEEE Internet of Things Journal*, vol. 10, no. 5, pp. 1234–1245, 2023.

[6] A. Kumar, "Ctrl movie review: Ananya panday is in control in this timely lesson on the dangers of ai," *The Hindu*.

[7] M. Constantinescu, "Genai avatar judges and virtuous adjudication," *Preprint*.

[8] H.-S. N. M. Voznak, "A bibliometric analysis of technology in digital health: Exploring health metaverse and visualizing emerging healthcare management trends," in *Proceedings of the IEEE Conference on [Conference Name]*, pp. 1–10, IEEE, 2023.

[9] S. Middaugh, "Calling nurse avatar," *Johns Hopkins Nursing Magazine*.

[10] H. Ismail, "Covid-19 vaccine misinformation aspects," 2025.

[11] R. Emmott, "Russia deploying coronavirus disinformation to sow panic in west, eu document says," *Reuters*.

[12] P. Gaborit, "A sociopolitical approach to disinformation and ai: Concerns, responses and challenges," *Journal of Political Science and International Relations*, vol. 7, no. 4, pp. 75–88, 2024.

[13] B. Group, "Trusted tv doctors 'deepfaked' to promote health scams on social media," *BMJ Group*, 2024.

[14] E. Hayward, "Deepfakes of michael mosley used to sell health scams," *The Times*, 2024.

[15] M. R. Mehra, S. S. Desai, F. Ruschitzka, and A. N. Patel, "Hydroxychloroquine or chloroquine with or without a macrolide for treatment of covid-19: A multinational registry analysis," *The Lancet*, vol. 395, no. 10240, pp. 1820–1826, 2020. Retracted on 4 June 2020 due to data inconsistencies.

[16] C. Piller and K. Servick, "Two elite medical journals retract coronavirus papers over data integrity questions," *Science*, 2020.

[17] A. Johnson, B. Lee, and C. Martinez, “Advancements in renewable energy forecasting using machine learning techniques,” *IEEE Transactions on Sustainable Energy*, vol. 14, no. 3, pp. 456–467, 2023.

[18] E. Davis and M. Brown, “Ethical implications of ai-generated content in journalism,” in *Proceedings of the 2024 ACM Symposium on AI Ethics*, pp. 98–105, ACM, 2024.

[19] R. Wilson and D. Thompson, “Critical review of data, models, and performance metrics for wind and solar power forecast,” *IEEE Transactions on Sustainable Energy*, vol. 12, no. 1, pp. 789–798, 2021.

[20] V. Heath, “Can we stop ai fuelling the spread of climate change denial and misinformation?,” *Geographical*.

[21] C. Staff, “No, florida won’t be hit by a category 6 hurricane as viral tiktok video claims,” *ClickOrlando*, 2023.

[22] O. SAF, “Debunking false claims about sea ice,” *EUMETSAT*, 2022.

[23] J. Doe, J. Smith, and A. Johnson, “Ai threats to politics, elections, and democracy: A blockchain-based deepfake authenticity verification framework,” *Journal of Artificial Intelligence Research*, vol. 2, no. 4, pp. 123–145, 2024.

[24] J. Thomas and D. Chazan, “Deepfake videos of macron posted online — by the president himself,” *The Times*, 2025.

[25] B. Rieder and M. van Kleek, “Beyond porn and discreditation: Epistemic promises and perils of deepfake technology in digital lifeworlds,” *Media and Communication*, vol. 8, no. 3, pp. 1–4, 2020.

[26] P. Contributors, “Hong kong protestors implement methods to avoid facial recognition technology and government tracking,” 2022.

[27] A. N. Staff, “New hampshire primary biden ai deepfake robocall sparks concern over election integrity,” *Associated Press*, 2024.

[28] M. Meaker, “Slovakia’s election deepfakes show ai is a danger to democracy,” *WIRED*, 2023.

[29] P. Trifonova and S. Venkatagiri, “Misinformation, fraud, and stereotyping: Towards a typology of harm caused by deepfakes,” in *Proceedings of the 2024 ACM Conference on Computer-Supported Cooperative Work (CSCW)*, pp. 1–9, ACM, 2024.

[30] N. C. Köbis, B. Doležalová, and I. Soraperra, “Fooled twice: People cannot detect deepfakes but think they can,” *iScience*, vol. 24, p. 103364, Nov. 2021.

[31] M. Dupuis, K. Renaud, and R. Searle, “Crowdsourcing quality concerns: An examination of amazon’s mechanical turk,” in *The 23rd Annual Conference on Information Technology Education*, (Chicago IL USA), p. 127–129, ACM, Sept. 2022.

[32] M. Dupuis, B. Endicott-Popovsky, and R. Crossler, “An analysis of the use of amazon’s mechanical turk for survey research in the cloud,” in *International Conference on Cloud Security Management*, (Seattle, Washington), Oct. 2013.

[33] P. G. Ipeirotis, F. Provost, and J. Wang, “Quality management on amazon mechanical turk,” in *Proceedings of the ACM SIGKDD Workshop on Human Computation*, (Washington DC), p. 64–67, ACM, 2010.

[34] M. Groh, Z. Epstein, C. Firestone, and R. Picard, “Deepfake detection by human crowds, machines, and machine-informed crowds,” *Proceedings of the National Academy of Sciences*, vol. 119, p. e2110013119, Jan. 2022.

[35] D. Chesney, Bobby; Citron, “Deep fakes: A looming challenge for privacy,” 2019.

[36] C. Vaccari and A. Chadwick, “Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news,” *Social Media + Society*, vol. 6, p. 2056305120903408, Jan. 2020.

[37] A. J. Nederhof, “Methods of coping with social desirability bias: A review,” *European Journal of Social Psychology*, vol. 15, no. 3, p. 263–280, 1985.

[38] C. L. Kimberlin and A. G. Winterstein, “Validity and reliability of measurement instruments used in research,” *American Journal of Health-System Pharmacy*, vol. 65, p. 2276–2284, Dec. 2008.

[39] P. M. Podsakoff, S. B. MacKenzie, J.-Y. Lee, and N. P. Podsakoff, “Common method biases in behavioral research: a critical review of the literature and recommended remedies.,” *Journal of Applied Psychology*, vol. 88, no. 5, p. 879–903, 2003.

[40] S. B. MacKenzie and P. M. Podsakoff, “Common method bias in marketing: Causes, mechanisms, and procedural remedies,” *Journal of Retailing*, vol. 88, no. 4, p. 542–555, 2012.

[41] J. D. West and C. T. Bergstrom, “Misinformation in and about science,” *Proceedings of the National Academy of Sciences*, vol. 118, p. e1912444117, Apr. 2021.

[42] D. G. Johnson and N. Diakopoulos, “What to do about deepfakes,” *Communications of the ACM*, vol. 64, p. 33–35, Mar. 2021.

[43] Á. Vizoso, M. Vaz-Álvarez, and X. López-García, “Fighting deepfakes: Media and internet giants’ converging and diverging strategies against hi-tech misinformation,” *Media and Communication*, vol. 9, no. 1, pp. 291–300, 2021.