CENTER FOR NATIONAL SECURITY STUDIES

EMBARGOED CONTENT: NOT FOR PUBLIC RELEASE PRIOR TO OCT 28TH 2024 AT 10:30 AM MST.

Research Summary: Impact of AI Generated Media

Executive Summary

AI generated deepfakes are a growing concern with regard to election security and disinformation. Deepfake content has already been used in attempts to influence elections or geopolitics several times all over the globe, including in Utah. This problem will increase in severity as generative AI improves, and it becomes more difficult for voters to discern between authentic content and deepfakes. A cross-disciplinary coalition of UVU faculty, staff, and students has conducted a study to measure the impact of deepfake content versus real media, monitor viewers' non-conscious responses to deepfakes, and ascertain how well participants can identify deepfakes retrospectively.

Key takeaways from this study include:

1. Deepfakes tested in this study received equivalent or higher ratings than authentic content in categories including credibility, trustworthiness, and persuasiveness.
2. Participants had difficulty discerning whether the content they had viewed was real or AI generated. Over 50% of participants rated deepfake content as 'probably real' or 'definitely real.'
3. Biometric data showed higher levels of engagement and confusion when exposed to deepfake content, as evidenced by micro-expressions, though they did not report these feelings during post-test interviews. This suggests that deepfakes may trigger a non-conscious response associated with the "uncanny valley" effect.

Project Background and Structure

The increasing sophistication of AI-generated deepfakes poses a serious threat to election security and public trust in information. As generative AI continues to improve, distinguishing authentic content from deepfakes will become even more difficult for voters. If deepfake content becomes indistinguishable from real media, the impact on public trust and election security could be severe.

To address this, the Center for National Security Studies and the Herbert Institute are partnering with UVU's Neuromarketing SMART Lab. This state-of-the-art lab, known for its work in neuroscience and marketing, utilizes biometric technology like eye-tracking, facial coding, galvanic skin response, and EEG to assess unconscious responses to digital content. For this study, we are leveraging these tools to evaluate how deepfake content compares to authentic media in terms of effectiveness and credibility.

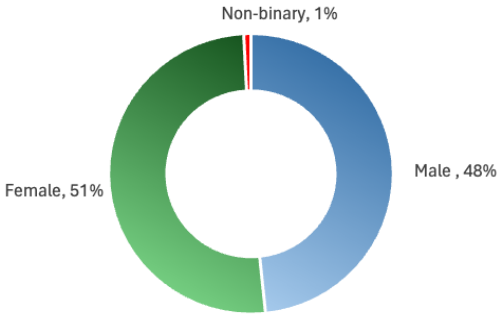
CENTER FOR NATIONAL SECURITY STUDIES

Study Design

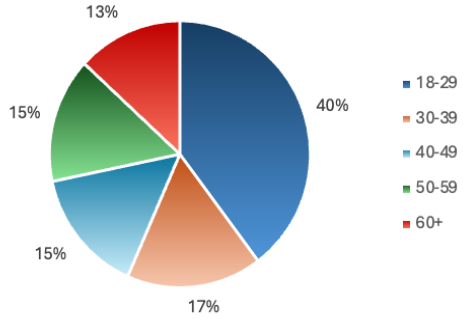
Unlike other studies, we aimed to replicate the natural conditions under which people encounter deepfake disinformation. Specifically, we wanted subjects to interact with short-form media samples while in a natural state of mind in which they are not aware that they may be being deceived. Accordingly, participants were not subjected to heightened stress or alertness while viewing the media. Students from the Center for National Security Studies created deepfake and control content, including video and audio samples. We controlled for factors such as timing, messaging, and demographics to ensure the integrity of our data.

Demographics

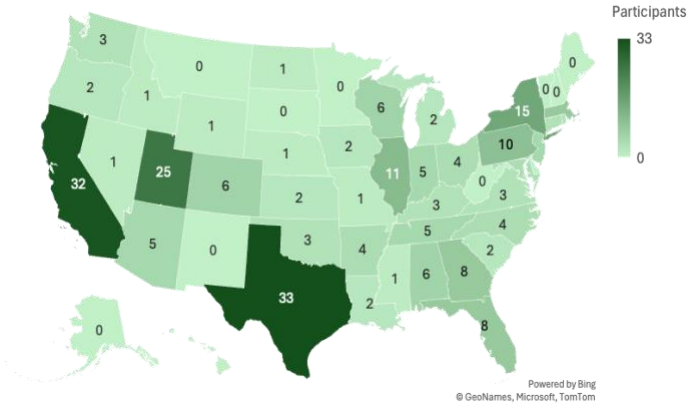
Gender



Age



State Distribution



To minimize bias, we selected non-controversial topics, ensuring participants' preconceived opinions did not bias the results or increase their awareness during the study. This study is designed to be the first in a series, providing a foundation for future research on more focused topics. Potential follow-up studies could explore the impact of deepfakes on down-ballot

elections or examine whether people are more susceptible to deepfake content that aligns with their existing beliefs.

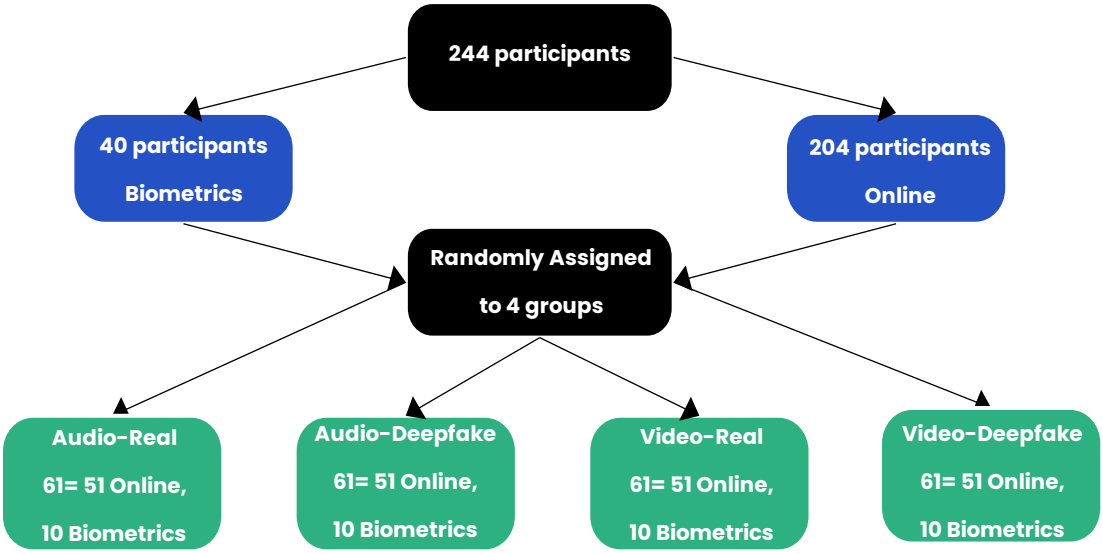
Research Questions

We designed this study to address four key questions:

1. Is there a measurable difference in the credibility of legitimate media versus deepfake media?
2. Do participants exhibit different unconscious responses to real versus deepfake content?
3. How accurately can subjects identify deepfake media after viewing or listening to it?
4. Is there a difference in the ability to distinguish deepfakes in audio versus video content?

Methodology

A total of 244 subjects participated in the study, with 40 of them tested on-site to collect biometric data, including eye-tracking and facial coding. The participants were divided into four equal groups and exposed to either a video or audio sample.



At the beginning of the test, participants were unaware that some content was AI-generated. After viewing or listening to the media, participants evaluated the message and speaker on factors such as credibility, knowledge, and trustworthiness. Participants would rate the content they viewed on a Likert scale (1-7) with one being the least favorable rating, four being neutral, 7 being the most favorable. They were then given the opportunity to explain their rating in a short-answer response. Questions in this section of study were as follows:

1. What was your impression of the speaker? (Short Answer)
2. How knowledgeable do you think the speaker is about the topic? (Likert Score & Short Answer)
3. How trustworthy do you think the speaker is about the topic? (Likert Score & Short Answer)
4. How persuasive do you find the speaker? (Likert Score & Short Answer)
5. How reliable did you find the information in the sample? (Likert Score & Short Answer)
6. How would you rate the overall quality of the content? (Likert Score & Short Answer)
7. This content seemed authentic. (Likert Score & Short Answer)

Following this section, subjects were informed that the study aimed to measure the impact of deepfakes, and that some content may have been AI-generated. Participants were then asked to assess whether they believed the media was real or AI-generated and to rate their confidence in their judgment.

Results

- **Impact on Viewer**:
  Deepfake and genuine media were rated by participants across several categories, including the speaker's knowledgeability, trustworthiness, persuasiveness, reliability of the information, and quality of the content. The average ratings across each the categories showed that deepfakes had effectively the same impact on viewers as real content. No statistically significant differences were observed between deepfake content and real media. Meaning that in effect, deepfake content was just as impactful as real content.

Likert Scale

1	2	3	4	5	6	7
Least Favorable	Highly Unfavorable	Somewhat Unfavorable	Neutral	Somewhat Favorable	Highly Favorable	Most Favorable

CENTER FOR NATIONAL SECURITY STUDIES

Perceived Knowledge			
Video		Audio	
Real	Deepfake	Real	Deepfake
5.59	5.95	5.77	5.64

Perceived Reliability			
Video		Audio	
Real	Deepfake	Real	Deepfake
5.44	5.44	5.57	5.44

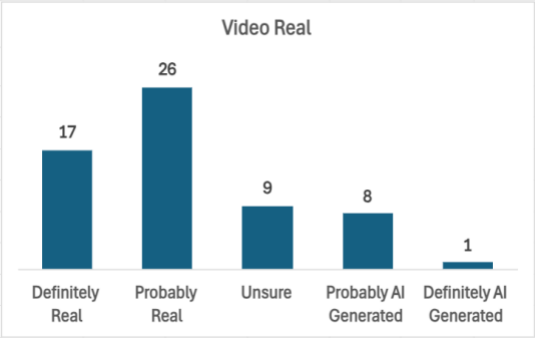
Perceived Trustworthiness			
Video		Audio	
Real	Deepfake	Real	Deepfake
5.61	5.62	5.7	5.64

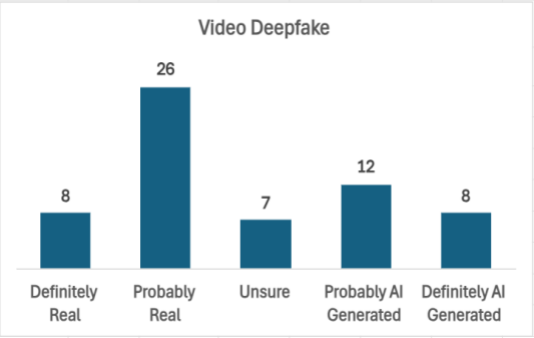
Content Quality			
Video		Audio	
Real	Deepfake	Real	Deepfake
5.21	5.64	5.97	6.02

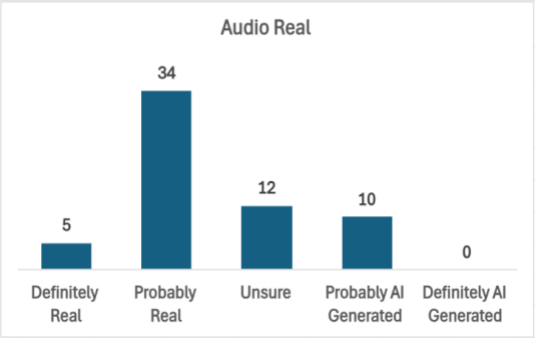
Perceived Persuasiveness			
Video		Audio	
Real	Deepfake	Real	Deepfake
5.26	5.41	5.48	5.57

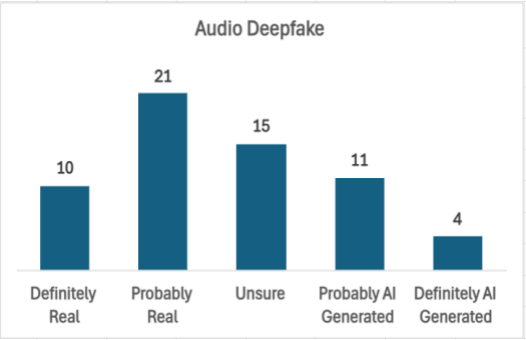
Perceived Authenticity			
Video		Audio	
Real	Deepfake	Real	Deepfake
5.41	5.25	5.9	5.56

- **Difficulty Identifying Deepfakes in Retrospect:**
  Even after being informed that they might have encountered a deepfake, participants struggled to consistently identify AI-generated content. Across all media types—real video, deepfake video, real audio, and deepfake audio—at least 50% of participants believed the media was "probably real." Furthermore, 57% or more were confident in their assessment. This suggests participants had at best, a 50% chance of detecting a deepfake, with most maintaining their original judgments even after learning that AI-generated content might have been included.









- **Non-conscious Engagement with Deepfakes:**
  Participants showed higher levels of engagement and confusion when exposed to deepfake content, as evidenced by micro-expressions, though they did not report these feelings during post-test interviews. This suggests that deepfakes may trigger a non-conscious response associated with the "uncanny valley" effect. In contrast, real media prompted more traditional emotional responses which were also expressed more strongly than emotions elicited by deepfakes

Confusion	
DF Video	**1.93**
DF Audio	**1.20**
Real Audio	**1.00**
Real Video	**0.74**

Engagement	
DF Video	**11.82**
Real Video	**10.81**
DF Audio	**10.11**
Real Audio	**2.90**