**Exploring the Impact of Deepfake Warnings on People's Video Perception**

Venkatanaveenkumar P

The Pennsylvania State University,   
Researcher -2

The Pennsylvania State University,

Researcher-3

The Pennsylvania State University,

Deepfake videos are hyper-realistic videos generated by artificial intelligence to portray someone saying or doing things that never actually happened. Research shows that implementing warning labels reduces the deepfake misinformation in social media. However, current research indicates that the presence of deepfake warning labels does not significantly improve viewers' ability to identify deepfakes, and the cognitive impact of warning labels on viewers has yet to be studied. In our study, we assess the attention shifts and impact of warning labels on the viewers while watching videos and aim to provide insights for designing more effective warning labels through the analysis of the results. Through a mixed-method approach which consists of thematic analysis and quantitative analysis, we analyzed the psychological and cognitive responses of participants to interview videos. We conducted interviews with participants showing both real and deepfake videos to analyze the attention patterns and perceptions across two phases 1) Phase 1- without deepfake warning label and 2) with deepfake warning label. Our findings reveal that the warning labels can enhance participants' awareness of deepfakes, and higher awareness leads participants to disperse their attention towards different visual elements to identify the deepfake videos. At the same time, the effectiveness of warning labels changes based on the individual attitude and prior knowledge of deepfake. Our study can offer insights into designing effective warning labels.

CCS CONCEPTS • Deepfake Warning Label • Human Attention Shifts • Warning Label Impact

ACM Reference Format:

Lingwei Qu, Shubhom Rawat and Venkatanaveenkumar Prabhuleti. 2024. Exploring the Impact of Deepfake Warnings on People's Video Perception. In Proceedings of the 2024 Final Project Paper IST 577: Human Factors of Security and Privacy, The Pennsylvania State University, University Park, December 11, 2024.

1. Introduction

With the fast development of Artificial Intelligence (AI) models, deepfake technology has advanced to the point of producing hyper-realistic yet influenced content. Deepfakes are the product of artificial intelligence (AI) applications that merge, combine, replace, and superimpose images and video clips to create fake videos that appear authentic [1]. Coupled with the reach and speed of social media, convincing deepfakes can quickly reach millions of people and have negative impacts on our society [9]. In such cases, people's ability to discern misinformation is concerning. As demonstrated by earlier studies, people have a hard time correctly recognizing information created by AI [2, 5]. This has already become a global issue, as across all countries and all media types, people rated AI-generated samples as more likely to be produced by a human than a machine [3].

To combat the growing threat of deepfakes, in addition to developing more effective deepfake detection models, efforts are also being made to help end users discern whether a video is a deepfake on their own. Adding warning labels to deepfake content on social media is one such approach. According to research, adding corrective labels can help reduce belief and sharing of misinformation [10, 11]. However, most of these studies focus on deepfake images and text content. When it comes to deepfake videos, warning labels appear to be less effective: even with the presence of warning labels, the vast majority of individuals are still unable to spot a deepfake from a genuine video even when they are told the content they are viewing may have been altered [9].

Building on these studies, our research aims to investigate how would deepfake warning labels affects people’s perception and attention while watching videos. We seek to bridge the knowledge gap regarding users' visual and cognitive engagement with deepfakes by integrating participant-driven feedback with quantitative focus area data. We conducted a study with six participants, asking them to watch two sets of videos, each consisting of one real video and one deepfake video, and provide feedback. Before showing the second set of videos, we presented a deepfake warning label. We then analyzed the differences in participants' feedback on the two sets of videos to determine the impact of the deepfake warning label on their attention distribution while watching the videos.

We obtained several findings. First, the presence of a deepfake warning label indeed shift people's attention, leading them to focus on details within the video rather than the overall content, and using the quality of those details to judge whether a video is a deepfake or not. Second, the influence of the warning label lasts for a period of time and is not limited to the labeled video. Lastly, individuals' awareness of and attitudes toward deepfakes affect the effectiveness of the warning label.

By understanding the cognitive impact of warning labels on individuals, our research can assist in designing more effective warning labels, thereby further reducing the belief in and spread of misinformation.

1. RELATED WORK
   1. Typical Attention Pattern

If we want to study how warning labels impact people's attention allocation, we first need to understand where attention typically allocates during normal interactions between individuals. According to related research, consistent and genuine eye contact can help maintain trust and engagement in conversation [12]. At the same time, audiences are naturally attuned to eye contact, which can influence their attention during presentations [13]. Apart from eye contact, studies have also shown that emotional facial expressions tend to capture attention [14].

We can infer that when people do not need to worry about misinformation caused by deepfakes, their attention is likely to focus on the speaker's face, carefully observing their expressions, and focus on the spoken content.

* 1. Deepfake and Warning Label Research

Frank et al. (2024) performed a comprehensive study across multiple countries to examine the ability to detect various AI-generated media. Their findings showed that humans perform poorly in identifying the real and AI-generated media by state-of-the-art methods, with an average accuracy of just 54%, which is nearly like 50-50 probability and never crossed 60% for other media types. This was due to the influence of cognitive and personal variables such as generalized trust, cognitive reflection, familiarity with deepfakes, and holistic thinking [3].

Hishami et al. (2024) analyzed human perception of audiovisual deepfakes through a gamified web-based platform. Their research distinguishes between human observers and state-of-the-art deepfake detection models. Their study revealed that humans tend to overestimate their detection capabilities, suggesting a potential overconfidence bias in assessing the authenticity of digital content [4].

Lewis et al. (2023) explored the effectiveness of content warnings in enhancing deepfake detection. Their research study provided that while warning labels increased participants’ suspicion towards content, they did not drastically improve the detection accuracy of deepfakes. Their findings suggest that effective identification and mitigation of deepfake content may require trust in external authentication sources and enhanced content warnings because general content warning labels are not capable of enhancing human detection capabilities [5].

* 1. Codebook Analysis

Ortloff et al. (2023) provided valuable insights into the qualitative analysis process, especially in the context of security advice research. Their study mentioned that a minimum of two coders is required to analyze complex data in qualitative analysis, such as interview transcripts, benefiting from interactions between coders while refining and creating the final codebook [6]. Halpin (2024) offered guidance on calculating and interpreting inter-coder agreements in qualitative coding. This study highlighted the importance of Cohen’s Kappa when analyzing the reliability between coders, providing a more efficient measure of agreement that accounts for chance agreement [7]. The related work highlights the challenges in human detection of deepfakes. While humans struggle with detecting deepfakes and AI models outperform in media generation and detection, there remains a significant gap in understanding human attention shifts, cognitive factors, and the effectiveness of human judgment in detecting deepfakes.

1. Research questions and STUDY DESIGN
   1. Research Questions

For conducting our study, we have developed two research questions.

**RQ1: How's user's attentional focus differ between video labeled with deepfake warning and unlabeled videos?**

The question explores the direct impact of warning labels on the viewer’s attention. It is crucial to determine whether such labels direct the user’s attention away from the video’s substance and onto a particular aspect, including facial expression or odd details shown in the video. By Defining this topic, the study can focus on the direct impact of warnings on perception, providing information about how well these labels work as a mitigating measure.

**RQ2: Does the presence of a 'deepfake warning' label on one video influence viewers' attentional focus on another, unlabeled video?**

The other question investigates if viewer’s perception of later unlabeled videos is affected after being exposed to a deepfake warning label. It examines the possibility of spillover, in which the doubt sparked by one warning label affects how people pay attention to other things. The study asses the wider perspective of warnings on general media trust without significant provoking audio clip so that the viewer’s judgment remains unbiased by approaching this as a distinct subject.

While framing the experiment we have designed two hypothesis that we were expecting from our participants:

**Hypothesis I:** when people are watching videos without any deepfake warning labels, they usually pay more attention to the character’s spoken content. However, after receiving a deepfake warning, they start focusing on other aspects of video like unnatural facial expressions, lip movement, bad texture of background or visual artifacts that make them believe abnormalities or suggest a deepfake.

**Hypothesis II:** When a deepfake warning label is presented to the viewers, it affects how they perceive and interpret the later unlabeled videos as a whole. Following a warning label, viewers may demonstrate a similar attentional shift to that observed when viewing an actual deepfake content, and they may begin to doubt even the legitimacy of real, unlabeled videos.

* 1. Study Design

This section is the outline of how we have conducted our interview process to collect data from the participants.

* + 1. Interview Procedure

**Phase I**

*Step-1*: In the first step, when the participants are ready, they will be shown a set of videos randomly selected from three prepared video sets for Phase I. Each set includes one real video and one deepfake video, with the real video presented first followed by the deepfake video. The real one is shown first to have a baseline perception for the viewers, followed by the deepfake video for comparison. During the video playback, participants are free to rewind and pause the video at any instance if they want to rewatch any part again. After watching the video, we asked participants to circle the focused areas on printed screenshots of the two videos. By collecting the data of circled screenshot, we get the insight about their area of focus in the phase without label videos.

*Step-2*: Completing step-1, participants will be asked a set of open-ended questions to assess their comprehension and opinion about both videos. The questions were designed to assess if participants paid attention to the spoken content of the video, what visual elements they noticed, and their thoughts and feelings about the video. Please refer to the appendix for the complete list of interview questions.

**Phase II**

*Step-1*: In the second phase, the participants will again be exposed to a set of videos randomly selected from three prepared video sets for Phase II. Before presenting the videos to the participants, we will provide the following instructions to the participant: the video that is going to play next will have a warning label “The video might be generated by AI”. After making sure the content of the warning label is fully registered with the participant, we will proceed to the video. In Phase II, each video set consists of a deepfake video with a warning label, and a real video without a warning label. The deepfake video with a warning label will be presented first, followed by the real video without a warning. With this fixed order, we aim to check the impact of the label and the change of attention from Phase I that might be overlooked. After watching the video, we asked participants to circle the focused areas on printed screenshots of the two videos.

*Step-2*: The same set of questions from Phase I will be asked again in Phase II. There will be additional questions introduced to the participants, to examine whether there is any change in their attentional focus, and if they also evaluated the real video with questionable qualities, having the label in mind that video may be AI generated.

The interviews were conducted in person. We used Zoom to record and transcribe the interviews. We didn’t record the faces of participants due to privacy concerns. Interviews were conducted on average of eleven minutes and uploaded saved recordings to our Microsoft Teams project channel.

1. METHOD
   1. Deepfake Videos

We selected both real and deepfake videos from common social media platforms such as YouTube and TikTok for our study due to the following key factors: 1) video diversity, featuring various speakers from both genders and multiple races, 2) deepfake video with minor imperfections that are not easily noticeable without proper attention, 3) only one speaker talking with a fixed camera angle, and 4) the use of English as a common language, aligning with our research focus to ensure comprehension of the content spoken by the speakers.

We selected a total of six sets of videos, each containing two videos. We carefully analyzed, discussed, and divided the videos into each set to balance the cognitive load on participants, ensuring they were not overwhelmed if presented with particularly challenging deepfake videos. This approach helped minimize the loss of information on their attention shifts when answering interview questions in both phases. Please refer to the appendix for the sources of the videos.

* 1. Interview Study
     1. Analysis

We followed the Mixed-Method approach for data analysis. Mixed-Method approach: 1) Qualitative Analysis- With thematic analysis, we explored how deepfake warnings influence participants’ psychological responses and perceptions of videos. We categorized the data into different themes based on the questions we asked, and the responses given. Each theme consists of relevant codes labeled and defined on the basis of themes, questions, responses, and comprehensiveness towards videos. These codes were applied to the interview data to get the participants’ attention. 2) Quantitative Analysis - We calculated the percentage weights of themes from the coding process (in the 4.2.2 section), i.e., from the participants’ responses.

* + 1. Codebook Generation

Codebook creation consists of a dedicated procedure and rules. Transcribed responses were thoroughly checked and corrected for any spelling mistakes made by the Zoom algorithm. The refined transcription was converted into an Excel chart. This chart included all the questions asked in both phases as one main column, with the other columns containing participants’ responses to each question in a row. This process made it easier to draw meaningful insights for creating the codebook.

A diagram of a codebook

Description automatically generated

Figure 1: Codebook Refinement and Theme Weight Calculation Process in Interview analysis.

We are three researchers conducting a research study on human attention toward deepfake videos. Two of our researchers served as coders and started coding (labeling the characteristics observed in the study) based on their individual knowledge. A minimum of two coders is required to create a codebook to mitigate individual perceptions, opinions, and biases in the complex data collected during the interview process [6]. While creating the codebook, the coders refined their codes by referring to the original transcription chart. Also, the two coders discuss each other on their individual codebooks after finishing their own codebooks. During this time, coders will go through a lot of arguments and agreements on their codes and definitions, also interpreting the participants’ responses. After finishing all the discussions, they came up with a refined codebook and next discussed with the third coder to avoid bias towards the interview data. It will create a new refined and well-organized final codebook for further analysis (see Figure 1).

The final codebook was created by defining six themes: Content Elements, Visual Elements, Judging Authenticity, Cognitive and Emotional Reactions, Impact of Warning Labels, and Attitude Toward Deepfake. Each theme is designed to capture specific characteristics of human attention, cognition, and comprehension related to the videos (See in Figure 13).

* + 1. Code Process

After creating the final refined codebook, we counted how many participants referred to the content of each code by using the codebook as a reference. To minimize bias and misinterpretation during this process, each researcher was responsible for counting codes for their two interviewed participants as well as two participants interviewed by another researcher on the team. This approach ensured an unbiased count. In cases of disagreement between researchers, they discussed the discrepancies with each other and involved the third researcher to finalize the correct count for a particular code. This process happens with all the three researchers and the count will happen twice for each participants’ response. This will produce a final code count for all participants for each code. We then calculated the weights of each code under each theme to find the difference in participants’ attention shifts. Next, we averaged weights for each theme to better understand the participants’ focus and thought processes while watching the videos.

* + 1. Inter-Coder Agreement Score.

Two coders analyzed the whole data independently at first then developed a codebook through iterative discussions under the supervision of third review coders. The inter-coder agreement by using Cohen’s Kappa is 0.91. This statistical score refers to agreement between coders on a set of categorical items, where it shows 0.91 which is a good indication of high consistency in their evaluations and the agreement is strong.

* 1. Participants Recruitment

For our study, we required participants open to all ages, gender, and academic background. We conducted interviews with a total of six participants, with each researcher responsible for recruiting and interviewing two individuals. The interview was conducted with proper consent from the participants. Participants are from different academic backgrounds, ages and across different geographical regions (Please refer to Table 1).

Table 1: Demographics and academic backgrounds of participants

|  |  |  |
| --- | --- | --- |
| Item | Option | Percentage |
| Gender | Male  Female | 66.67%  33.33% |
| Age | 18 – 24  25 – 34  55 – 64 | 50%  16.67%  33.33% |
| Academic Background | Cybersecurity – P1, P2  Housewife – P3  Professor in Business – P4  Computer Science – P5, P6 | |

1. RESULT

In this section, we will conduct an analysis based on the interview data and attention heatmap data.

* 1. Analysis of Interview Data

Based on the codes we compiled from the interview data, we categorized them into six major themes: **Content Elements**, **Visual Elements**, **Judging Authenticity**, **Cognitive and Emotional Reactions**, **Impact of Warning Labels**, and **Attitude Toward Deepfake**. Then we aligned them with our three main topics that we mainly focusing on: **Attention Shift**, **Perception of the Videos**, and **Impact of the Warning Label**. However, one of the themes (**Attitude Toward Deepfake**) does not fall under the predefined topics, as we had not expected it to arise.

* + 1. Attention shift

In the experiment, to examine whether the presence of a deepfake warning label would shift people's attention during watching the videos due to, we designed the following questions.

**Q1: Please briefly summarize the videos.**

**Q3: Do you feel yourself focusing on any particular thing in the video?**

By comparing participants' responses to the same questions after watching the two sets of videos, we aimed to assess whether they demonstrate an attentional shift.

Under this topic, we identified two related themes: **Content Elements** and **Visual Elements**. The first theme includes two codes: **content\_focus** and **context\_clarity**. The second theme is composed of seven codes: **face**, **facial\_expressions**, **eye\_movement**, **lip\_movements**, **background**, **peripheral\_elements**, and **whole\_screen**. Please refer to the appendix for the complete codebook.

Table 2: Weight Changes of Theme “Content Elements” and “Visual Elements” Across Different Phases

|  |  |  |
| --- | --- | --- |
| Theme | Phase I | Phase II |
| Content Elements | 66.7% | 75.0% |
| Visual Elements | 42.8% | 35.7% |

From Table 1, we can observe that the weight of content elements increased in Phase II compared to Phase I, while the weight of visual elements decreased. Based solely on the theme weight data, this result appears to contradict our hypothesis. However, upon analyzing the specific content of the interviews, we identified the following phenomena:

The increase in weight was caused by the individual abilities of two participants. **P6** showed no attention to content in either phase (which means he neither mentioned focusing on the spoken content, nor accurately summarized the video content). This might be due to an oversight by the researcher conducting the interview, who failed to prompt P6 to summarize the spoken content after P6 omitted doing so. As a result, P6 did not recount any specific spoken content from the videos.

**P3**, on the other hand, initially confirmed her readiness to begin the session, but underestimated the experiment's cognitive load. When the video started, she failed to process it effectively and forgot that she was allowed to rewatch the video, resulting in retaining none of its content. In Phase II, however, P3 was better prepared for the cognitive demands of the experiment and was aware that she needed to summarize the video content. As a result, she made an extra effort to memorize it.

If the data from these two participants were excluded, we would find that the remaining four participants all demonstrated attention to the spoken content of the videos in both phases and were able to summarize the content with relative accuracy.

However, during our detailed analysis of the responses, we found that all four participants provided longer descriptions of the video content in Phase I compared to Phase II. For example, when testing the same video set, **P2** and **P4** provided the following content summaries in Phase I regarding the real video:

P2: The first video is a 13-second video in which a man tells his experience about football, like kind of trauma in which we he had concussions while playing football due to the all the ramming between the players and all.

P4: The first one was about concussions and playing football and boxing. I think it was a boxer, but he was talking about football originally and then getting hit with helmets to helmets playing football, and how that can lead to concussions. And he started that a young kid and looked to me like he was playing later in life, you know, more than a young kid. And I inferred he had some problems from that.

In contrast, their summaries of the real video in Phase II were as follows:

P2: And then the second video that's very real. It's from an interview of Obama. So, in that thing, he was just explaining something.

P4: Second video was something it looked to me like it was about jobs for young people...... I thought he was trying to say something important, but the full video message didn't quite get out. I don't know exactly what his point was, but something about, I think, jobs for young people are related, waiting for something.

From their responses, it is evident that their summaries in Phase II became noticeably shorter compared to Phase I and leaned more toward providing an overall topic summary. P2 failed even to summarize the topic, while P4 believed that the spoken content in the video was incomplete and failed to convey the speaker's viewpoint effectively. However, the video contained a full and coherent statement of the speaker's viewpoint (“I think with young people, you don’t always need to be impatient, asking for the plum assignment.” [Group 6, real video]). This indicates that in Phase II, the attention these two participants allocated to the spoken content was reduced compared to Phase I.

It is worth noting that although P2 failed in his attempt to summarize the genuine video content, he successfully identified the core topic of the deepfake video. Additionally, he was confident and quick to conclude that he was watching a deepfake video. This suggests that once P2 confirmed he had successfully identified the deepfake video, he was able to reallocate his attention to the spoken content.

In the analysis of the codes for visual elements, we observed shifts in the distribution of attention across different elements. The areas participants focused on varied depending on the video, indicating that the attention patterns in Phase I were disrupted. Initially, attention was relatively concentrated on certain specific elements, but it became dispersed across other elements. Where the attention shifted depended on where participants believed the "flaws" in the video were based on their observations, resulting in the loss of a relatively unified pattern.

A diagram of a diagram

Description automatically generated

Figure 2: Weight Shift of Each Code within Theme Visual Elements Between 2 Phases.

In Phase I, more participants focused their attention on the face, facial expressions, and the background revealed due to the speaker's movements in the video, which suggests a pattern where, without a deepfake warning label, participants tend to focus on visual elements such as the face as a whole and facial expressions to understand the underlying message. Their attention then naturally shifts to other elements that capture their interest (“The next one was talking about that kind of magic, and getting you to focus your attention on the coin, and looked like he was trying to do some magic to confuse you. So I was looking to see if there was something else going on that it's going to redirect my attention” (P4, Phase I, judging Group 3 deepfake video)). However, the proportions of these elements decreased in Phase II. Instead, more participants shifted their attention to peripheral elements, the whole screen, as well as lip movements and eye movements (see Figure 2).

This indicates that the overall decrease in the weight of visual elements was due to the previously concentrated attention being drawn to specific elements, which were often those that participants perceived as suspicious indicators of a deepfake (e.g., “But in the second one, the lip, the lip movement seemed a bit unnatural and also the right shoulder, it seemed quite blurred” (**P5**, Phase II, judging Group 3 deepfake video), “In the first video, the camera angle, the expressions and overall, the moment of the body was very rigid, and the expressions weren't changing that much” (**P6**, Phase II, judging Group 4 deepfake video)).

**Finding 1**: *Our thematic analysis confirmed our hypothesis: the presence of the deepfake warning label will cause an attention shift. Participants tend to distribute their attention across different visual elements in the video, attempting to identify suspicious elements. They then focused more attention on these areas compared to the spoken content, to determine whether they could serve as indicators of a deepfake (****RQ1****)*.

* + 1. Perception of the Video

In order to examine whether the presence of a deepfake warning label would have an impact on how people perceive the videos, we designed the following questions:

**Q2: Can you describe the overall experience after watching the video?**

**Q4: What are your thoughts on the two videos?**

By asking these relatively broad, open-ended questions without a clear directional focus for responses, and then comparing the differences in answers before and after the appearance of the deepfake warning label, we aim to investigate whether the perception of videos by participants changes due to the presence of the label.

Under this topic, we identified two related themes: Judging Authenticity and Cognitive & Emotional Reaction. The first theme includes six codes: **authenticity\_visuals**, **credibility\_topic**, **suspected\_deepfake**, **deepfake\_knowledge, content\_knowledge,** and **speaker\_knowledge**. The second theme is composed of two codes: **emotional\_response,** and **cogntive\_effort**. Please refer to the appendix for the complete codebook.

Table 3: Weight Changes of Theme “Judging Authenticity” and “Cognitive & Emotional Reaction” Across Different Phases

|  |  |  |
| --- | --- | --- |
| Theme | Phase I | Phase II |
| Judging Authenticity | 44.5% | 38.8% |
| Cognitive & Emotional Reaction | 75.0% | 58.3% |

From Table 1, we can observe that the weight of **Judging Authenticity** decreased in Phase II compared to Phase I, while the weight of **Cognitive & Emotional Reaction** increased. Based solely on the theme weight data, the result from theme **Cognitive & Emotional Reaction** appears to align with our hypothesis, while the result from theme **Judging Authenticity** does not.

When analyzing specific interview content, we found that three participants (**P2**, **P5**, and **P6**) expressed suspicion that the video might be a deepfake as early as Phase I (see Figure 3). This could be attributed to their majors in information technology-related fields, which may make them more sensitive to issues related to deepfakes.

A group of black text

Description automatically generated with medium confidence

Figure 3: Code Pattern for Participants’ Who Tend to Judge Authenticity Before the Warning Label.

At the same time, an analysis of code weight reveals a pattern: the originally dispersed code weight becomes concentrated on specific codes, leading to an overall decrease in the theme's weight (see Figure 4).

A graph of red and blue bars

Description automatically generated

Figure 4: Weight Shift of Each Code within Theme Judging Authenticity Between 2 Phases.

This occurs because fewer codes are mentioned by a larger number of participants, while the number of participants referencing the remaining half of the codes drops significantly.

From this pattern, we can reasonably infer that the warning label plays a role in raising awareness of potential deepfakes (while **P2**, **P5**, and **P6** already demonstrated a certain level of awareness in Phase I), and they will tend to judge authenticity of the video because of the awareness. Additionally, people tend to rely on visual information to identify deepfakes, as instances of making judgments based on knowledge of the speaker and spoken content have decreased.

The other theme under this topic, Cognitive & Emotional Reaction, also supports this inference through its decreased weight. Before the deepfake warning label was introduced, when participants were asked about their thoughts and impressions of the video, their responses often reflected their emotional reactions to the video and the topic: “This seems like a clip from an interview. So, I will definitely find and watch the interview again to know all his trauma...... The second one is Tom Cruise performing a magic trick. So it was good to watch, but it is a very normal magic trick.” (P2, Phase I), “The first one seems like a social issue, like, the people playing with the football and injuring their heads...... The second one was more about entertaining and, you know, it was a good-looking guy. He kind of captured the smile of his face and he tried to, you know, that was kind of more playful.” (**P4**, Phase I).

In contrast, when asked the same question in Phase II, their responses changed to the following: “My overall experience is like, I got to know how bad people can do deepfake by seeing the first one, and that's like, yeah, this is like, one of the defects I have seen in a while with that bad editing” (**P2**, Phase II), “The overall experience of the first one I thought was fake and the second one I thought was real.” (**P4**, Phase II).

It is worth noting that P2 was identified as someone who exhibited sensitivity to deepfakes as early as Phase I (“Yeah, it comes a little on it because, at initial glance I thought it was a deepfake” (**P2**, Phase I, when asked why did he focus on those particular elements in the video he had paid attention to)), but he still did not mention his suspicions when asked about his overall experience, instead focusing more on his reaction to the video. After the deepfake warning label was displayed, his response immediately aligned with that of **P4**, who had not exhibited deepfake awareness in Phase I, as both completely omitted any mention of their emotional reactions to the topic and focused solely on their opinions about whether the videos were deepfakes.

**Finding 2**: *Our thematic analysis suggests that the presence of a deepfake warning makes individuals more inclined to judge whether the video is a deepfake or not, rather than engaging with or reflecting on their personal experience of the video's content and emotional impact (****RQ1****).*

* + 1. Impact of the Warning Label

To obtain more direct responses from participants, we designed the following questions to explore whether participants believed the presence of a deepfake warning label influenced them and whether this influence extended to videos without a deepfake label:

**Q3: Do you feel yourself focusing on any particular thing in the video?**

**Q5: Did the presence of a warning label change your perception of the videos?**

Under this topic, we identified one related theme: **Impact of Warning Label**. The theme includes two codes: **warning\_impact**, and **attentional\_shift**. Please refer to the appendix for the complete codebook.

We also identified an interesting pattern under this theme: all participants who felt the label influenced them reported that it caused an attentional shift (see Figure 5). When watching the second genuine video in Phase II (without the warning label), they approached it with the assumption that the first video was a deepfake, viewing the second video through the lens of "could this also be a deepfake?" As mentioned in the previous section, even though this was a genuine video, the attention patterns observed differed from those in Phase I, where there was no label prompt.

A black and white image of letters and numbers

Description automatically generated

Figure 5: Code Pattern for Participants’ Who Felt There Was an Impact of the Warning Label.

In contrast, participants who did not feel that the label influenced their perception of videos displayed nearly identical attention patterns in Phase II and Phase I (" Well, I was looking at the video as a whole ", “Again, I was looking at the video as a whole” (P5, Phase I and Phase II)). We will discuss in the next section why they believed the label had no impact on them.

**Finding 3**: *Our thematic analysis confirmed our hypothesis: The presence of a warning label raises participants' level of awareness, specifically prompting them to think, "I need to pay attention to whether this might be a deepfake or not." This heightened awareness leads to a pattern of attention distribution that differs from how their attention is distributed when watching videos without a warning label (****RQ2****).*

* + 1. Attitude Toward Deepfake

We asked participants who believed the label had no impact on them to explain their reasoning. Among them, P1 and P3 shared relatively similar views: both felt that the content of the videos was not particularly important and that it wouldn’t matter much even if the videos were AI-generated.

**P1:** No, I actually forgot the warning. because I was listening to them...... I don't think it (the label) would affect me. it was not really, like, unbelievable. It felt like someone was telling the real things. So, it was not like they're making up stuff and talking. I honestly forgot that the label exists.

**P3**: No. I didn't feel anything different. It did not because they were not talking about anything controversial...... No, if it's just saying something about, pretty much about nothing. I don't think it's provoking any dislikeness in me.

From their responses, we can see that **P1** believed the spoken content did not resemble misinformation, so she did not bother to determine whether the video was a deepfake. **P3**, on the other hand, felt that the spoken content was not controversial, making it unnecessary to put extra effort into discerning the deepfake clue.

**P3**'s perspective might be influenced by her age and the type of media she regularly interacts with. She does not frequently engage with social media and is not particularly interested in current events, which could shape her attitude toward deepfakes. In contrast, **P4**, who belongs to the same age group with **P3**, demonstrates greater awareness and sensitivity to current events, likely due to his profession. As a professor, he interacts with a wide range of people in his daily life, and his teaching role requires him to take greater responsibility for the information he shares and the opinions he expresses. It is possible that this greater responsibility makes him more attentive to deepfakes as a potential source of misinformation.

As for **P1**, we were surprised that, despite her background in cybersecurity, she exhibited a different attitude compared to other students from information and computer technology-related disciplines. However, after conducting a detailed analysis of her responses, we developed some possible hypotheses.

**P1:** The first 2 videos I saw were easy to take along. I was able to understand what the context was, maybe because of the familiarity with the characters on the video, I could relate that, OK, they are from the movie industry and they're talking something about the movies. But because I don't know any of these 2 people here (in Phase II videos), I had to take in time to understand what the context starts because he can literally talk about anything...... And also, they didn't layout a pretext or an introduction for what they're speaking.

We observed that P1 prioritized “fully understanding the spoken content of the video” over “assessing the reliability of the information source”, which in this context translates to “determining whether the video is a deepfake”. Given that the videos used in our experiment were only 10 to 15 seconds long, she was unable to fully grasp the complete message conveyed in the videos. Since the higher-priority task (understanding the spoken content) was not yet completed, the lower-priority task (evaluating whether the video was a deepfake) did not proceed.

This is further evidenced by the fact that P1 was the only participant who rewound a video: during Phase II, she replayed the real video because she had no idea what it was about in the first viewing. Additionally, when asked about her overall experience of the videos in Phase II, her evaluation of the deepfake video was as follows: “And the first one was fine, because it gave an introduction. But if there were no introduction for people who don't know gravity or relativity, it wouldn't be helpful”. We can see that she places great importance on the idea that "the video content should be understandable". Therefore, her attitude toward deepfakes may be related to the fact that her priorities when evaluating information during its reception differ from those of other participants.

As for P5, we found that, even in Phase I without any prompts, he immediately began assessing the video's authenticity. He would first evaluate the video as a whole and then quickly focus on the parts he deemed suspicious. This pattern remained consistent across both phases:

**P5** (Phase I): Well, I was looking at the video as a whole. I found that the first video was quite clear, like it had defined textures and stuff, but in the second one, it was a bit blurred, and part of the videos had fuzzy details...... The lip, the lip movement seemed a bit unnatural, and also the right-hand shoulder seemed quite blurred.

**P5** (Phase II): Again, I was looking at the video as a whole. I mean, I noticed that the textures in the first video were kind of abnormal. In the first video, there were no physical features of the face, not clearly defined as compared to that in the second video.

It is likely that his attention pattern while watching videos had already been shaped by his heightened vigilance toward deepfakes. As a result, he perceived the warning label as unnecessary, believing that he could confidently and easily identify a deepfake even without the label's prompt (“But if the warning would not have been there, I think I would have deemed that to be a deep fake because of the texture and the quality of the video” (P5, Phase II when asked if he thinks the warning label changed his perception of the video)). Compared to P1, it is clear that P5 prioritized determining whether a video was a deepfake over fully understanding its content, as in Phase II, P5 was unable to accurately summarize the spoken content of any video.

**Finding 4**: *Attitudes toward deepfakes can also influence the effectiveness of the warning label. Whether participants believe that deepfakes do not affect the authenticity of the content or that identifying a deepfake is sufficient regardless of the content's importance, both perspectives can invalidate the warning label: in such cases, their attention patterns remain unchanged regardless of whether a video might be a deepfake.*

* 1. Analysis of Attention Heatmap Data

During the interviews, to verify the consistency between subjective and objective data, we asked all participants to mark areas in video screenshots where they paid attention to. In this section, we will analyze the heatmap data obtained from this process.

* + 1. Group 1

This group of videos was tested on **P1** and **P6** in Phase I.



Figure 6: Group 1 Video Attention Heatmap. Tested in Phase I. (a) is the real video, and (b) is the deepfake.

It is important to note that P1 prioritized understanding the video content, while P6 exhibited awareness toward deepfakes even without the label. In this set of videos (see Figure 6), we observed that P1’s attention pattern remained consistent across both videos, focusing primarily on the speaker's face. In contrast, P6 paid attention not only to the face but also to the background and the clothing folds under the speaker's arms.

Interestingly, P6 did not mention noticing the underarm area during the interview. However, when the researcher asked him to confirm whether he had marked that dot accidentally, P6 clarified that he had indeed paid attention to it. While the reason for not mentioning it during the interview is unclear, it is possible that he did not find anything unusual there. This further suggests that P6 might have been actively scanning the video for unnatural elements to determine whether it was a deepfake.

* + 1. Group 2

This group of videos was tested on **P3** and **P5** in Phase I.



Figure 7: Group 2 Video Attention Heatmap. Tested in Phase I. (a) is the real video, and (b) is the deepfake.

Similar to the videos in Group 1, P3 and P5 were also participants with distinct attitudes toward deepfakes. The difference lies in their perspectives: P3 believed that deepfakes were unimportant, while P5 demonstrated a high level of awareness regarding deepfakes.

In this set of videos (see Figure 7), we observed that P3’s attention consistently focused on the faces of the speakers in the video, while P5 also noticed the blurred reflections on the speaker's shoulders. This further supports the idea that P5 was actively searching for evidence that the video might be a deepfake even without the label.

* + 1. Group 3

This group of videos was tested on **P2** and **P4** in Phase I.



Figure 8: Group 3 Video Attention Heatmap. Tested in Phase I. (a) is the real video, and (b) is the deepfake.

Interestingly, all three groups of participants consisted of one individual without prior deepfake awareness and another with a strong sense of deepfake awareness, even if the assignment of video groups to participants was determined randomly. In this set of videos (see Figure 8), **P4**, as a participant who did not initially exhibit deepfake awareness, focused on the speaker’s face in the first real video. In contrast, **P2**, who demonstrated deepfake awareness, paid additional attention to the microphone (“I was focusing on this little mic. I was thinking where is the receiver?” (**P2**)).

Similarly, in the second video, P4’s attention was entirely guided by the content of the video. Since the speaker attempted a magic trick, **P4** shifted their focus to the props, trying to figure out how the trick worked (“...... I was looking for some kind of misdirection” (**P4**)). Instead of focusing on the entire face, **P4** concentrated on the mouth, as they found the smile particularly stood out (“...... it was a good-looking guy. He kind of captured the smile of his face......” (**P4**)). On the other hand, **P2**’s attention was not guided by the video’s content; instead, they focused on background elements (“I was least concentrated on the coin, because I know there was something behind him. May be a blue frame hanging, and I was trying to read it” (**P2**)). This divergence suggests that P2, leveraging their knowledge of deepfakes, was actively searching for unnatural aspects in the video, consistent with their pattern of behavior.

* + 1. Group 4

This group of videos was tested on P1 and P6 in Phase II.

A person with a yellow face

Description automatically generated

Figure 9: Group 4 Video Attention Heatmap. Tested in Phase II. (a) is the deepfake, and (b) is the real video

In this set of videos (see Figure 9), **P1** focused on the speaker's face in the deepfake video but stated that it wasn't intentional, while in the real video, she only noticed the ring, as she realized she couldn't understand the spoken content: “I wouldn't say I was focusing more on the face. I was also listening to him, so I was just looking at the looking at the screen..... I focused on his ring because I couldn't focus on what he was saying”.

**P6** focused on the speaker's face in both videos because he relied on facial expressions to determine whether a video was a deepfake: “...... In the first video, the camera angle, I mean the expressions and the overall moment of the body was very rigid and the expressions weren't changing that much. While in the second video there was a lot of changing of expression as well as the hand moment.”

* + 1. Group 5

This group of videos was tested on P3 and P5 in Phase II.



Figure 10: Group 5 Video Attention Heatmap. Tested in Phase II. (a) is the deepfake, and (b) is the real video.

In this set of videos (see Figure 10), **P5** claimed that he focused on the entire screen, yet during the marking process, he identified specific elements rather than circling the whole screen. This aligns with the attention pattern we summarized for him: because he is accustomed to first assessing whether a video is a deepfake, he tends to scan the screen for suspicious elements and then concentrate his attention on those areas.

His focus on the speaker’s face can be attributed to his tendency to evaluate the texture of facial skin as a criterion for determining whether a video is a deepfake (“In the first video, there were no physical features of the face. They were not clearly defined as compared to that in the second video” (**P5**)). While for the real video, he did not explicitly mention paying attention to the headphones during the interview, it is possible that he initially aimed to assess whether the texture of the headphones appeared unnatural but ultimately found nothing noteworthy.

* + 1. Group 6

This group of videos was tested on **P2** and **P4** in Phase II.

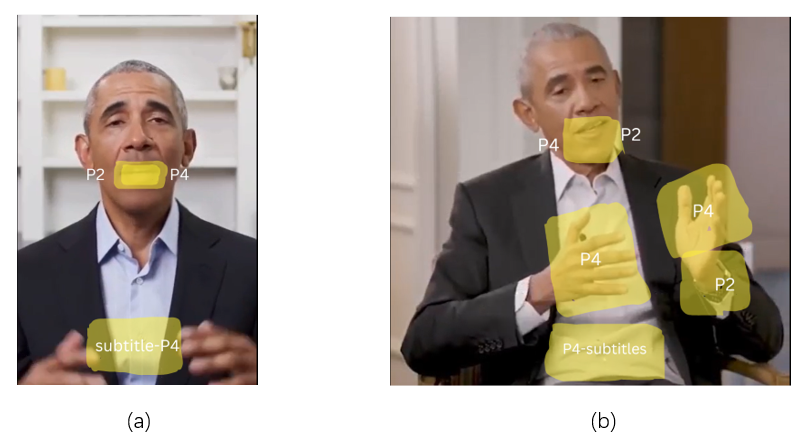


Figure 11: Group 6 Video Attention Heatmap. Tested in Phase II. (a) is the deepfake, and (b) is the real video.

In this set of videos (see Figure 11), we observed that when watching the deepfake video, both P2 and P4 focused on the speaker's mouth movements, as this was a notably "unnatural element." In addition, P4 also paid attention to the subtitles, which might reflect his habitual way of watching videos.

In the real video, both participants continued to focus on the speaker's mouth rather than their face. This pattern was carried over from the previous deepfake video, where they identified the deepfake by observing the speaker's mouth. This suggests that the effect of the label extended to the second video, even though it did not have a warning label.

Meanwhile, their attention patterns evolved further. P4 continued to focus on the subtitles and, due to his familiarity with the speaker, he observed the speaker's gestures to assess whether they aligned with his prior understanding of the speaker. P2, on the other hand, directed his attention to the speaker's watch, which might indicate that he was still searching for other unnatural elements, beyond the mouth, to determine if the video was a deepfake.

**Finding 5**: *Based on the analysis of the heatmap data, we can preliminarily conclude that awareness of deepfakes influences people's attention patterns when watching videos. Higher awareness leads to a greater focus on unnatural elements rather than the video as a whole. Additionally, the warning label has a more noticeable impact on individuals who perceive a stronger connection between deepfakes and the video content by heightening their awareness, as opposed to those who view the two as separate and unrelated.*

1. Limitations

Although we made our best efforts, this study still has several limitations. First, the study involved only six participants, which means that any outlier among them could significantly impact the overall results. For example, P3, who was entirely inattentive during Phase I and unable to recall any details, provided little to no usable information. In future work, increasing the number of participants would significantly enhance the reliability and generalizability of this study.

Secondly, a disproportionate number of our participants (4 out of 6) had backgrounds related to information and computer technology. Due to the limited social networks of some researchers, it was challenging to find more suitable interviewees. Compared to participants without a related background, these participants exhibited noticeably higher awareness of deepfakes, which likely affected our ability to collect sufficient generalizable data in Phase I regarding how people typically perceive videos under normal circumstances. In future work, we should not only increase the number of participants but also improve the diversity of their backgrounds to mitigate the influence of participants' professional knowledge.

Thirdly, some researchers, due to a lack of experience, were unable to strictly follow the interview protocol, resulting in missing interview data. Given the small number of participants, any missing data points can significantly affect the overall validity of the findings. In future work, researchers should become more familiar with the protocol and adhere to it more strictly to collect more comprehensive and valid data.

Fourthly, the content of the videos may also influence participants' performance, such as the number of suspicious elements, the complexity of the visuals, and whether the video guides the viewer's attention. Although researchers attempted to group videos based on their content, differences between videos still impacted the reliability of the quantitative data. In future work, we could consider generating deepfake videos that are content-consistent with real videos, allowing for greater control over variables.

Last but not least, due to limitations in time and the technical expertise of our team, we are unable to establish ground truth for this project. Ground truth is the basis for performance analysis in computer vision and image processing, as it serves as a benchmark to evaluate the accuracy of results, helping researchers to validate whether participants' judgments about deepfake videos are accurate or influenced by biases [6]. Without it, we lack a standard for comparison, making it difficult to measure how well participants’ evaluations align with an established baseline. It could also limit the generalizability of our findings, as the results may appear context-dependent, and it would be harder to make a comparison with other deepfake research. In future work, we should incorporate ground truth using state-of-the-art models, to enhance the credibility and generalizability of our findings.

1. Conclusion

Our study shows that people’s perception of videos and attention patterns change based on their level of awareness of deepfakes when watching a video. Individuals with higher awareness tend to allocate some of their attention to identifying unnatural elements in the video to determine whether it is a deepfake or not. As a result, the level of attention devoted to the video content decreased accordingly, due to the distraction of potential suspicious elements. The deepfake warning label can effectively enhance people’s awareness of deepfakes, and this awareness is sustained -- one label influences their perception of multiple videos within a short time frame. However, the label's effectiveness diminishes for individuals whose awareness is either too high or too low, as the impact of the label is less significant in these cases.

Based on these findings, we suggest that future deepfake warning label designs could include a brief summary of the video content to reduce users’ cognitive load when assessing whether a video is a deepfake. By providing a summary, users could grasp the spoken content more easily and focus on distinguishing the video's authenticity. Additionally, incorporating prompts such as “deepfakes may contribute to misinformation” could enhance awareness of the potential harm caused by deepfakes, further heightening their awareness. At the same time, it would be helpful to display the label throughout the video instead of only showing it at the beginning, so viewers can maintain awareness continuously.

REFERENCES

1. Maras, M. H., & Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. International Journal of Evidence & Proof, 23(3): 255–262.
2. Mai, K. T., Bray, S., Davies, T., & Griffin, L. D. (2023). Warning: Humans cannot reliably detect speech deepfakes. Plos one, 18(8), e0285333
3. Frank, J., Herbert, F., Ricker, J., Schönherr, L., Eisenhofer, T., Fischer, A., ... & Holz, T. (2024, May). A representative study on human detection of artificially generated media across countries. In 2024 IEEE Symposium on Security and Privacy (SP) (pp. 55-73). IEEE.
4. Hashmi, A., Shahzad, S. A., Lin, C. W., Tsao, Y., & Wang, H. M. (2024). Unmasking Illusions: Understanding Human Perception of Audiovisual Deepfakes. arXiv preprint arXiv:2405.04097.
5. Lewis, A., Vu, P., Duch, R. M., & Chowdhury, A. (2023). Deepfake detection with and without content warnings. Royal Society Open Science, 10(11), 231214.
6. Ortloff, A. M., Fassl, M., Ponticello, A., Martius, F., Mertens, A., Krombholz, K., & Smith, M. (2023, April). Different researchers, different results? analyzing the influence of researcher experience and data type during qualitative analysis of an interview and survey study on security advice. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1-21).
7. Halpin, S. N. (2024). Inter-Coder Agreement in Qualitative Coding: Considerations for its Use. American Journal of Qualitative Research, 8(3), 23-43.
8. Daniel Kondermann. 2013. Ground truth design principles: an overview. In Proceedings of the International Workshop on Video and Image Ground Truth in Computer Vision Applications (VIGTA '13). Association for Computing Machinery, New York, NY, USA, Article 5, 1–4.
9. Westerlund, M. (2019). The emergence of deepfake technology: A review. Technology innovation management review, 9(11).
10. Martel, C., & Rand, D. G. (2023). Misinformation warning labels are widely effective: A review of warning effects and their moderating features. Current Opinion in Psychology, 101710.
11. Ahmed, S. (2021). Fooled by the fakes: Cognitive differences in perceived claim accuracy and sharing intention of non-political deepfakes. Personality and Individual Differences, 182, 111074.
12. Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. Journal of Nonverbal Behavior, 24(4), 239–263.
13. Conty L, Tijus C, Hugueville L, Coelho E, George N. Searching for asymmetries in the detection of gaze contact versus averted gaze under different head views: a behavioural study. Spat Vis. 2006;19(6):529-45.
14. Vuilleumier, P. (2002). Facial expression and selective attention. Current Opinion in Psychiatry, 15(3), 291-300.

A  APPENDICES

A.1 Interview Questions

The questions asked in our interview study to participants.

A screenshot of a video test

Description automatically generated

Figure 12: Interview study questions.

A.2 Codebook

The final Codebook created by the coders.

A screenshot of a computer

Description automatically generated

Figure 13: Refined Final Codebook of our interview study

A.3 Video Sources

Real video:   
No 1 https://www.youtube.com/watch?v=DOypgVBMjwY

No 2 https://www.tiktok.com/t/ZP8L9ghVW/

No 3 https://www.tiktok.com/t/ZTYLJjsxH/

No 4 https://www.tiktok.com/t/ZTYL1EBMo/

No 5 https://www.tiktok.com/t/ZTYL1vQq1/

No 6 https://www.tiktok.com/t/ZTYLeQyBD/

Deepfake:

No.1 (Tom singing): https://www.tiktok.com/t/ZTY8bfoeU/

No.2 https://studio.infinity.ai/gallery

No.3 (Tom magic) https://www.tiktok.com/t/ZTY8tVrom/

No.4 https://studio.infinity.ai/gallery

No.5 https://studio.infinity.ai/gallery

No.6 (Obama): https://www.tiktok.com/t/ZTY8bqT5D/