Beyond the Naked Eye: Empirical Study of How People Perceive, Detect, and Respond to AI-Manipulated Videos

KANIZ FATIMA*, Rochester Institute of Technology, USA
Y. KELLY WU*, Rochester Institute of Technology, USA
ERSIN UZUN, Rochester Institute of Technology, USA

The growing presence of AI-manipulated videos presents a significant challenge to the integrity of online information. This paper presents findings from an empirical study with 490 participants in the United States to provide a holistic view of public engagement with this threat. We structure our analysis around three key areas: (1) how demographics and media habits influence general perceptions of prevalence; (2) the factors shaping detection accuracy, the calibration of confidence level, and the perceptual cues people rely on when viewing in-the-wild videos; and (3) the verification actions people take following suspicion. We find that while the public views AI-manipulated media as prevalent, participants struggled to distinguish authentic and AI-manipulated videos, often exhibiting poorly calibrated confidence. Furthermore, users rarely utilize available detection tools. These patterns highlight the insufficiency of human detection ability and the need of new approaches to enable improved user awareness, successful interventions, and effective mitigation.

Additional Key Words and Phrases: User Perception, Deepfakes, Trust in Digital Media, AI-Generated Videos, Behavioral Study

ACM Reference Format:
Kaniz Fatima, Y. Kelly Wu, and Ersin Uzun. 2026. Beyond the Naked Eye: Empirical Study of How People Perceive, Detect, and Respond to AI-Manipulated Videos. In *Proceedings of CHI Conference on Human Factors in Computing Systems (CHI '26).* ACM, New York, NY, USA, 23 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Advancements in artificial intelligence (AI) have made the manipulation and synthetic generation of videos easier and widely accessible. As the underlying AI models continue to advance, AI-manipulated videos[1] are becoming more difficult to differentiate from authentic ones. While fueling creativity and enriching entertainment opportunities, AI-manipulated videos are also increasingly used to mislead or deceive. Malicious uses of this technology are rapidly expanding including, but not limited to, non-consensual intimate imagery (NCII) [13, 60], interference in democratic processes [28, 53], fraud [38, 61], medical misinformation [11] and evidence tempering [4]. Recent reports highlight both the scale and acceleration of this threat: the regions of the world seeing annual increases ranging from 410% to 1720% in fraud powered by this technology [61, 62], more than $410M in financial damages within the first half of 2025

*Equal contribution.
[1]We use the term *AI-manipulated videos* to refer to videos that has been created, altered or modified using AI-based techniques (e.g., face-swapping, lip-syncing, reenactment). We avoid using the term *deepfake* in our work because of its inconsistent use in public discourse. We retain *deepfake* only in two contexts: (a) in our survey instrument, where it is accompanied by a clear definition to reflect a familiar vocabulary to participants, and (b) when referring to prior publications that explicitly use the term.

Authors' Contact Information: Kaniz Fatima, kf2366@rit.edu, Rochester Institute of Technology, Rochester, NY, USA; Y. Kelly Wu, kellywu@mail.rit.edu, Rochester Institute of Technology, Rochester, NY, USA; Ersin Uzun, ersin.uzun@rit.edu, Rochester Institute of Technology, Rochester, NY, USA.

Manuscript submitted to ACM

alone [50], and the number of publicly available tools to create AI-manipulated audio and videos recently surpassing 3,000 [53]. These numbers clearly emphasize the magnitude and the urgency of the problem.

At a time where seeing is no longer believing, we still lack a comprehensive understanding of how the general public perceive, detect, and respond to AI-manipulated videos, particularly when engaging with content that mirrors the types of media they encounter online. Prior work has discussed the potential individual, political, and institutional harms posed by AI-manipulated media [16, 25, 32], and empirical studies have examined specific aspects of the problem, such as reactions to it within particular communities [12, 54] or public and journalistic discourse surrounding it [6, 72]. Research on human detection ability has separately shown that people often struggle to distinguish authentic from manipulated content, though such work typically relies on constrained or lab-curated datasets [27, 59]. Taken together, these studies offer important but partial insights into the broader AI-manipulated media landscape, yet they do not reveal how the general public perceives and engages with such content or how such engagement varies across demographic groups, media habits, and other individual factors. These gaps motivate the following research questions:

- **RQ1:** How do people's demographics and media habits shape their general perceptions of the AI-manipulated media landscape?
- **RQ2:** What factors influence individual accuracy, how calibrated is their confidence, and what perceptual cues do people rely on in detecting AI-manipulated videos?
- **RQ3:** What actions do people take after encountering suspected AI-manipulated videos, and what is their awareness and usage of available detection tools[2]?

To answer these questions, we conducted a three-part survey study from May to June 2025 with 490 participants recruited to approximate a representative sample of the general U.S. adult population [7]. The study first captured demographic information from participants and their overall views on AI-manipulated media. We then evaluated their detection abilities through a task where they judged the authenticity of a diverse, in-the-wild corpus of 13 videos. Our focus on the video modality is motivated by its increasing popularity among AI-manipulated media and its inherently richer multi-modal context, which integrates both visual and audio signals. Finally, we gathered their self-reported evaluation strategies, trust behaviors, and awareness and usage of detection tools. By examining perceptions, performance, confidence, perceptual cues, and behaviors within a single unified framework, our study offers a holistic view of how people experience AI-manipulated videos. Our findings show that while participants recognize AI-manipulated media as a part of the current social media ecosystem, their ability to detect manipulated videos is modest and often accompanied by poorly calibrated confidence. Furthermore, our findings reveal a mix of reactions to suspected content, from active informal verifications to passive ignoring it, alongside low awareness and usage of specialized detection tools. These insights offer a foundational, empirically grounded understanding of how the public perceives and engages with AI-manipulated videos, helping to guide the creation of practical strategies tailored to users to enhance public resilience in an ever-changing digital environment.

2 Related Work

2.1 Public Perceptions of AI-Manipulated Media

Research on people's perceptions of AI-manipulated media often focuses on niche populations, limiting insight into the views of the general public. For example, studies have focused on professional groups like journalists [56] and intelligence analysts [70], or sampled limited populations such as university students [12]. Other work narrows its

[2] *Detection tool* in this paper refers to any automated system designed to identify whether a given input media has been AI-manipulated.

scope to particular applications, including synthetic social media profiles [35, 52] and NCII [5, 66]. To gauge broader public sentiment, other work has analyzed online discourse. Discussions on Reddit span topics from politics and NCII to detection methods and regulation [20], and analysis of YouTube comments showed negative sentiment toward deepfake technology [10]. Discourse analyses of popular media and journalistic narratives similarly highlights its societal risks, particularly threats to individuals and the erosion of shared notions of truth [6, 72]. Together, these lines of work suggest that concerns about AI-manipulated media are widespread. A critical consequence of this media landscape is the growing uncertainty about the authenticity of digital content. Studies have found that exposure to potential deepfakes and concerns about them are associated with decreased confidence in news on social media [2, 68]. Such dynamics also contribute to the "liar's dividend," wherein heightened skepticism makes it harder for audiences to distinguish genuine content from fabricated material [65]. From the technology-acceptance perspective, public perceptions of AI-manipulated media are likely to evolve; *experience* with AI technologies can also potentially shift how people perceive it [31].

Building on this foundation, our work shifted attention from discourse samples and specialized contexts to a broader public perspective. We surveyed a representative sample of the U.S. adult population to assess their perceived prevalence of AI-manipulated media on social media and common use of it. Specifically, we quantitatively examined which factors, such as demographic characteristics and media habits, relate to the public's perceived prevalence, offering insights into how these perceptions may be distributed.

2.2 Human Detection of AI-Manipulated Media

While automated detection technologies continue to advance, demonstrating high accuracy across datasets [1, 9, 48, 73, 74], the critical challenge remains the human capacity to identify AI-manipulated media in naturalistic, highly contextualized settings such as social media. Studies measuring raw human performance on deepfake detection indicate a low baseline accuracy, hovering just above chance at approximately 55% across modalities and around 57% for video deepfakes specifically [17]. This unreliability is additionally accompanied by the tendency toward overconfidence, where individuals' perceived detection ability often exceeds their actual performance [26, 58].

To better understand this challenge, prior research has mapped detection capability against various individual characteristics. Lewis et al. found that younger individuals and those aware of deepfake technology are more likely to identify anomalies [29]. Furthermore, performance is influenced by the identity of the actor in the videos, with viewers showing better accuracy for individuals from their own ethnic group [24]. Evaluations of authenticity can also be influenced by personal bias, such as conservatism and agreement with the video's content [63]. Regarding visual cues viewers rely on, studies have found that individuals tend to focus on visual cues in the background and on facial features like the eyes, forehead, lips, and cheeks [64]. When presented with face-swapped videos, viewers have identified artifacts such as blurriness, unnatural expressions, and inconsistent skin color [69].

Prior research has substantially advanced our understanding of human detection ability for AI-manipulated media, but at this time, much of this work relies on constrained forms of video stimuli. As summarized by Somoray et al. [59], many studies rely on lab-grade datasets such as DeepFake Detection Challenge (DFDC) [18] or FaceForensics++ [51]. However, these datasets lack the variability or contextual richness of real-world online media. Moreover, prior work often focuses on particular types of content, such as political videos [23, 63], celebrity subjects [29], or unfamiliar actors [22, 24, 27, 64, 69]. Our study addresses this gap in ecological validity by collecting a diverse, in-the-wild corpus of authentic and manipulated videos sourced from online platforms. Additionally, we examined media habits (e.g., time spent on social media, preferred news sources) and trust behaviors (e.g., trust in sources and belief-consistent

information) in relationship to detection accuracy, extending beyond the basic demographic variables typical in prior work.

2.3 Verification Practices and Tool Usage

Beyond an individual's detection ability, the actions they take—or fail to take—after encountering suspicious media are key indicators of their knowledge on available resources and their overall verification mindset.

Prior research has looked into verification behaviors across different populations. Shahid et al. [54] investigated low-resource communities in India and observed a low willingness to actively identify fake videos and selective sharing patterns, particularly when content aligned with participants' worldviews. In contrast, Sohrawardi et al. [57] focused on journalists and found that they tend to adhere to traditional journalistic practices of verifying the context of the media first. Together, these studies highlight important differences in verification norms across communities but leave open questions about how the broader public approaches suspicious media.

The rapid advancement of detection technology has introduced a new verification pathway: web-based detection tools such as DeepFake-o-Meter [40], TrueMedia [67], Deepware [14], and AI or Not [44], which allow users to upload media for automated analysis. While researchers have begun to explore the design and usability of detection tools [33, 57, 70], little is known about whether the general public is aware of these options or use them in their everyday verification practices.

Despite the growth of technological solutions designed to combat AI-manipulated media, we lack a comprehensive understanding of how the general public navigates this landscape. To address this gap, we examine both the verification actions people report taking when encountering suspicious videos and their awareness and use of available detection tools. This perspective allows us to characterize common verification practices and surface barriers and opportunities for more effective interventions.

3 Experimental Design

To answer the three research questions outlined in Section 1, we used an online survey to capture both quantitative and qualitative insights. We describe the detailed recruitment, survey instruments, and our data collection procedure below.

3.1 Recruitment and Participants

We recruited a total of 556 voluntary self-selecting participants using the Prolific platform [47]. To ensure a representative sample of the U.S. adult population, our sampling employed Prolific's built-in features to balance participants across age, gender, and ethnicity based on U.S. Census Bureau statistics [7]. The reward for completing the study was set at $5 per participant, targeting a median compensation rate of $14.50/hr (200% of the 2025 federal minimum wage [41]). The actual median compensation rate translated to $12.61/hr, due to slightly longer completion times than expected.

After the initial collection, we performed data quality checks (detailed in Sec 3.2.4) and removed 66 responses that were either incomplete, duplicates, failed attention checks, or did not meet response-time criteria. This process yielded a final sample of 490 responses for analysis. The demographics of these 490 participants are summarized in Fig.1. Among the final participants, 59 reported having a disability (Visual: 3, Hearing: 5, Cognitive: 6, Other, mostly identified as physical: 40, Prefer not to say: 5).

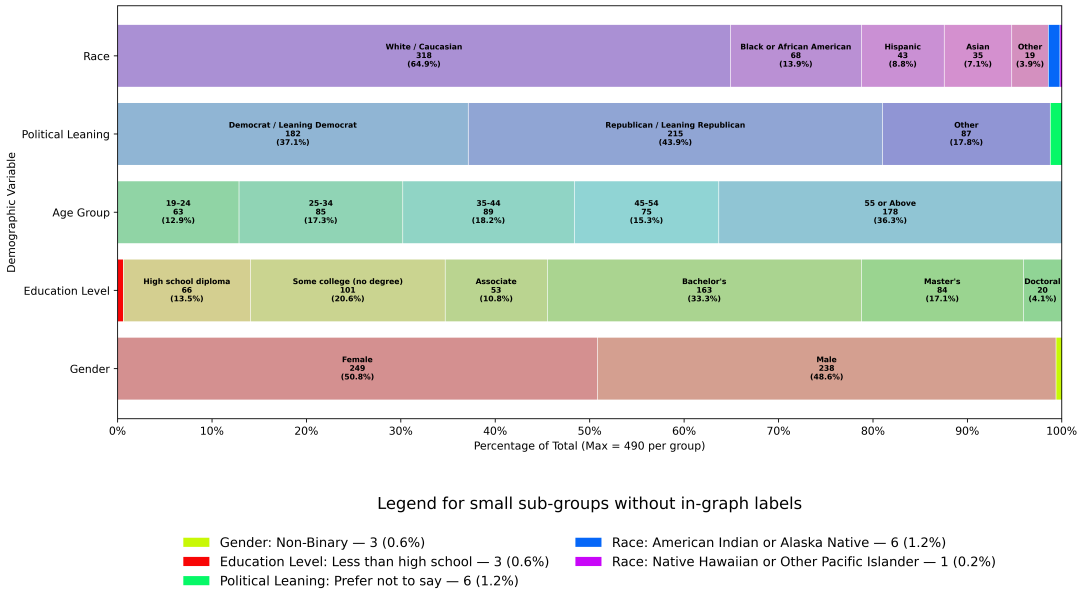


Fig. 1. Stacked bar chart showing the composition of the final sample of 490 participants across gender, education level, age group, political leaning, and race. The labels for sub-groups smaller than 2% of the sample are in the legend.

3.2 Survey Instrument

We used Qualtrics [49] for hosting the survey and data collection. The survey was open from May to June 2025. The online survey consisted of three logical parts outlined below. The exact questions used in each part can be found in Appendices A, B, and C.

3.2.1 Part 1: Demographics, media habits, and perceptions regarding AI-Manipulated Media. This first part of the survey was designed to collect data on participants' demographics (i.e., age, gender, education level, ethnicity, and disabilities), political leaning, general media habits (i.e., average time spent on social media [19] and ranking of news sources used [42]), and participants' perceptions of AI-manipulated media, including perceived prevalence on social media and common uses.

3.2.2 Part 2: Ability and confidence in identifying AI-Manipulated videos. The second part of the survey was designed to test participants' ability to correctly identify AI-manipulated and authentic videos, as well as to collect the self-reported confidence levels on a 5-point Likert scale and reasoning behind their selection in free-text responses. We selected 23 videos from a candidate pool of more than 200 public videos, which were sourced from online platforms such as YouTube, TikTok, and Facebook. Using videos already circulating online allowed us to construct a video set that better reflects the kinds of content people commonly encounter in everyday digital environments. The final video set (available at https://github.com/Anonymous-1a2b3c/Video-Selection) is summarized in Table 1.

In curating the final video set, we intentionally varied several factors that prior work suggests may influence perceived authenticity. For video-level factors, we considered overall visual quality, particularly resolution (1-A/1-B; 2-A/2-B) and camera stability (9-A/9-B), which has been identified as a factor worthy of further investigation [17]. We also varied the

presence or absence of source logos by cropping out broadcaster branding (6-A/6-B; 7-A/7-B), allowing us to examine the role of source cues as credibility indicators in online media environments [30]. We also varied content-level factors by selecting videos that naturally differed in the familiarity of the depicted person (e.g., public figures vs. unknown individuals) [17] and in the language spoken (English vs. non-English) [24, 37].

For the first 20 videos, we created 10 matched pairs (1-A/1-B through 10-A/10-B). Each pair differed primarily in one factor, which could be at video-level, content-level, or a case in which one video was AI-manipulated and the other was an authentic clip of the same person. Participants were randomly assigned to see only one video from each pair, enabling between-subjects comparisons. The remaining three videos (11, 12, and 13) were shown to all participants.

Video Number	AI-manipulated	Short Description	Length (secs)	Resolution	Participant Group Video Shown To
1-A	YES	Jennifer Lopez, Faceswap with synthetic audio	8	1920x1080	$1A$
1-B	YES	1-A in low resolution	8	256x144	$\neg 1A$
2-A	NO	Rachel McAdams	29	1920x1080	$2A$
2-B	NO	Rachel McAdams	29	256x144	$\neg 2A$
3-A	NO	Barrack Obama	30	1920 x 1080	$3A$
3-B	YES	Barrack Obama, lip-sync video with authentic audio from 3-A	30	1920 x 1080	$\neg 3A$
4-A	NO	Donald Trump	10	1920 x 1080	$4A$
4-B	YES	Donald Trump, Lip-sync with synthetic audio	10	1920 x 1080	$\neg 4A$
5-A	NO	Kim Jong Un	30	1920 x 1080	$5A$
5-B	YES	Kim Jong Un, Reenactment with synthetic audio	30	1920 x 1080	$\neg 5A$
6-A	NO	Jake Tapper	30	1920 x 1080	$6A$
6-B	NO	Jake Tapper, video in 6-A with CNN logo removed	30	1920 x 1080	$\neg 6A$
7-A	YES	Anderson Cooper, Lipsync with synthetic audio & post-processing	30	1920 x 1080	$7A$
7-B	YES	Anderson Cooper, video in 7-A with CNN logo removed	30	1920 x 1080	$\neg 7A$
8-A	NO	Matthew Miller	30	1920 x 1080	$8A$
8-B	YES	Matthew Miller, lipsync with synthetic audio on 8-A	30	1920 x 1080	$\neg 8A$
9-A	YES	Tom Cruise, faceswap with synthetic audio, distracting camera movements	30	1920 x 1080	$9A$
9-B	YES	Tom Cruise, faceswap with synthetic audio, stable camera position and audio	30	1920 x 1080	$\neg 9A$
10-A	NO	Unknown female actor #1, in Korean	30	1920 x 1080	$10A$
10-B	YES	Unknown female actor #2, faceswap with unknown audio source in Japanese	30	1920 x 1080	$\neg 10A$
11	YES	José Mourinho, faceswap with synthetic audio	30	1920 x 1080	*All*
12	YES	AI generated actor, synthetic audio and video	30	1920 x 1080	*All*
13	NO	Unknown female actor #3	30	1920 x 1080	*All*

Table 1. Short Descriptions of the videos used in the study. For participant groups: "*All*" represents all the participants; each other labeled group (e.g., "1*A*", "2*A*", etc.) represents a randomly selected independent subgroup of participants and includes approximately 50% of all the participants; "¬" sign before a subgroup label represents all other participants that are not in that particular subgroup

3.2.3 Part 3: Cues and behavioral responses. The last part of the survey collected self-reported data about the cues users notice first when encountering an AI-manipulated video, and how influential (5-point Likert Scale) various video specific cues (e.g., overall video quality, audio artifacts, audio/video synchronization issues, facial expressions, etc.) are in raising suspicion for them. Similarly, we asked them to rate their trust level for videos from various online sources,

including well-known media/news website and apps, social media posts by familiar or unfamiliar people, and their agreement level to statements testing whether an alignment with their existing views, knowledge, and beliefs is more influential than the reliability of the source in their decision making (both on a 5-point Likert Scale). Finally, we asked participants to select the typical actions they take after suspecting a video is AI-manipulated, and inquire whether, and if so which, tools they know of and use to check authenticity.

3.2.4 Quality Control Measures. For quality control, our survey included two attention check questions, one in Part 1 asking participants to choose the answer that starts with letter "F" in a multiple-choice question and another, in Part 3, checking whether they choose *"strongly disagree"* or *"disagree"* for the statement *"I swim across the pacific ocean to get to work **everyday**"* embedded in a question with multiple statements to rate. In addition, to minimize order effects bias [75], we randomized, for each participant, the order of the 13 videos, most multiple-choice options, and the statements in Likert-scale items. We did not randomize options with a meaningful inherent order (e.g., increasing time ranges), demographic items were presented in a fixed order, and in all multiple-choice questions the "Other" option was always placed last. The median time for participants to complete the survey was 23.8 minutes. For response-time check, we omitted the respondents that took less than 8 minutes to complete the survey as they were outliers.

3.3 Ethical Considerations

Our study design and protocol was reviewed and approved by our Institutional Review Board (IRB). We obtained informed consent from participants online and only those with explicit consent and attested to being at least 18 years old were allowed to take the survey. Throughout the study, we did not collect any personally identifiable information and the access to the collected data was limited to the IRB approved personnel with *Social & Behavioral Research Training* certification.

4 Results

4.1 RQ1: Perceptions of AI-Manipulated Media by Demographics and Media Habits

4.1.1 General Perceptions of AI-manipulated Media Prevalence and Application. Our findings indicate a universal belief that AI-manipulated media content has a notable presence on social media. When asked to estimate its prevalence, the majority of participants (63.7%, $N = 312$) believed it constituted between 6% and 30% of all content, with the 11-20% range being the most common selection ($N = 123$). However, opinions diverged at the extremes. A notable minority (8.2%, $N = 40$) perceived the presence of such media to be "Less than 5%". Conversely, another group (6.1%, $N = 30$) expressed a high level of concern, believing that AI-manipulated content makes up more than half of all media on social media. While exact figures on the prevalence of AI-manipulated media remain difficult to quantify, the concentration of estimates in the 6-30% range suggests that AI-manipulated media is no longer viewed as a niche occurrence, but an integral part of participants' daily social media feeds.

Regarding the application of this technology, a significant portion of participants (40.4%, $N = 198$) identified "political misinformation or election interference" as its most common use. The second most-selected application was "entertainment" (25.3%, $N = 124$), indicating that participants also recognize non-malicious uses. In contrast, "scientific or medical misinformation" was perceived as the least common application, selected by only 8 participants, which suggests that this emerging threat [11, 34] has not yet reached widespread public awareness. Underscoring the perceived societal impact, one participant who selected "Other" specified that the technology's main use is to *"stir up social pressures."*

4.1.2 Factors Associated with Perceived Prevalence. To investigate the factors influencing how prevalent participants believe AI-manipulated media to be, we conducted an ordinal logistic regression. The dependent variable was participants' estimates of AI-manipulated media prevalence, grouped into three ordered levels: low (≤ 10%), medium (11-30%), and high (>30%). The model included the following predictors[3]: gender (female vs. male), age (≤ 35 vs. >35), education (no college degree vs. college degree or above), political affiliation (Democrat or leaning-democrat, Republican or leaning-republican, other), daily social media usage (≤ 2 hours vs. >2 hours), primary news source (traditional vs. internet-based), and perceived use of AI-manipulated videos (entertainment vs. malicious). The proportional-odds assumption was assessed and found to be appropriate for the data.

The analysis results showed that demographic identity was a stronger predictor of perception than media habits. We found significant associations between gender and perceived prevalence: female participants had significantly higher odds of perceiving a greater prevalence compared to male participants ($Est. = 0.47, p = 0.006, OR = 1.60$). Participants identifying as Republican or leaning-republican had 58% higher odds of reporting a higher prevalence compared to their Democrat or leaning-democrat counterparts ($Est. = 0.46, p = 0.018, OR = 1.58$). Lastly, education showed a marginal positive association, with college degree holders showing a tendency to estimate higher prevalence ($Est. = 0.35, p = 0.056, OR = 1.41$), while age was not a significant factor.

In contrast, media habits did not show a statistically significant link to prevalence estimates. Our sample skewed towards heavy daily social media usage, with the largest group of participants (29.2%, $N = 143$) reporting they spend more than three hours per day on social media, while only a small fraction (4.1%, $N = 20$) reported spending less than 30 minutes. Despite social media being a common distribution platform for AI-manipulated content, daily social media usage was not a significant predictor in our model ($p = 0.34$). Regarding the news sources (see Fig. 2), more than half of our participants (50.6%, $N = 248$) indicated social media to be their primary news source, followed by mainstream news websites and apps (22.9%, $N = 112$); traditional news sources like TV or radio were only considered primary by 78 participants. However, we did not find a link between a participant's primary news source and their prevalence estimates ($p = 0.81$).

RQ1 Key Insights

AI-manipulated media is now perceived as an integral part of the online ecosystem. Most people accept AI-manipulated media as a common feature of social media.

Media consumption habits do not predict perception. Surprisingly, neither the amount of time spent on social media nor a person's primary news source had a significant association with their perception of how prevalent AI-manipulated media is. This suggests the perception formation around prevalence in this context is more complex than the simple exposure frequency.

Female participants perceived a greater prevalence. Gender in our study shows significant relationships with perceived prevalence of AI-manipulated media. This may be related to broader dynamics in which women are disproportionately targeted by malicious uses AI-manipulated videos and images [13], but this possibility should be further explored.

[3]For all regression analyses, we excluded participants who selected "Non-binary/Other" for gender (N=3) due to insufficient sample size, and the "Other" political category included participants who selected "Prefer not to say" (N=6).

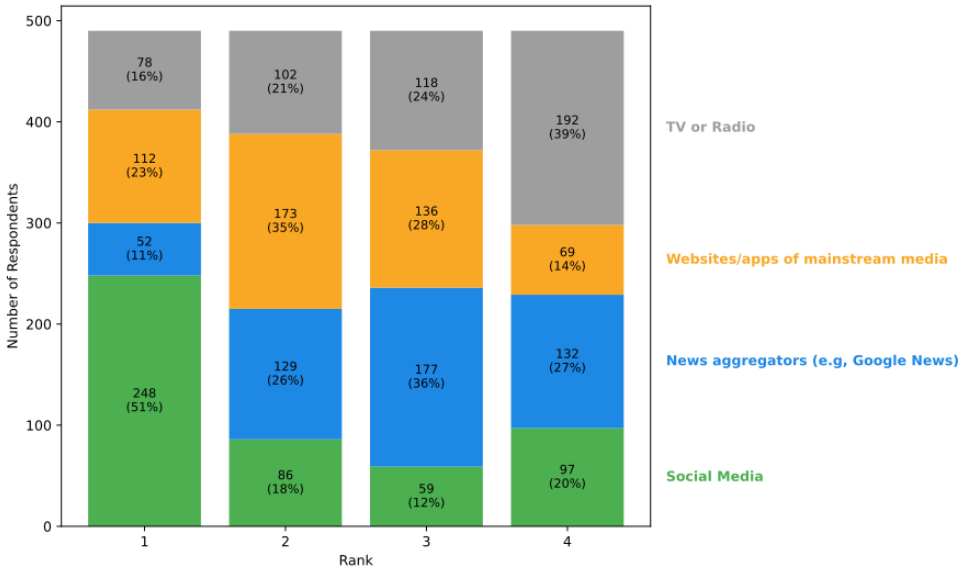


Fig. 2. Stacked bar chart showing how participants ranked their sources for news (from most common to least common). Percentages and counts are displayed within the stacked bars.

4.2 RQ2: Factors Influencing the Evaluation of Video Authenticity

4.2.1 Participant Accuracy and Confidence in Detecting AI-manipulated Videos. We began by establishing a baseline for participants' ability to distinguish authentic videos from manipulated ones. Across the 13 videos each participant was assigned to evaluate, the mean accuracy was 66.3% ($M = 8.61$ videos correctly identified, $SD = 1.87$), a rate better than chance but still highlights the challenge these videos pose. A notable pattern emerged when separating videos by their authenticity: participants were considerably more successful at identifying authentic videos (75.6% accuracy) than they were at identifying manipulated videos (59.4% accuracy). This suggests a potential bias towards accepting content as authentic.

In terms of participants' reported confidence level on their judgment, the overall mean was 3.68 on a 5-point scale with 5 being the most confident ($SD = 0.74, Min = 1.07, Max = 5.00$). A Student's t-test revealed a significant link between accuracy and confidence overall; participants were significantly more confident when their judgments were correct ($M = 3.75, SD = 1.08$) compared to when they were incorrect ($M = 3.54, SD = 1.10; p < .01$). However, if we look at authentic and AI-manipulated videos separately, confidence level was a strong indicator only for authentic videos. Student's t-tests showed that participants were significantly more confident when they correctly identified an authentic video ($M = 3.80, SD = 1.03$) than when they incorrectly flagged it as a manipulated one ($M = 3.25, SD = 1.15; p < .01$); there is no significant difference in confidence between correctly identifying an AI-manipulated video ($M = 3.71, SD = 1.13$) and incorrectly identifying it ($M = 3.66, SD = 1.06; p = 0.20$). This result suggests that participants only experienced a statistically significant drop in confidence when making the error of incorrectly labeling an authentic video as manipulated. In contrast, their confidence remained rather high when they were fooled by a manipulated video.

To further investigate this relationship, we analyzed the alignment between participants' confidence and their accuracy by identifying those who were "overconfident" (below-mean accuracy but above or equal to mean confidence level) and "underconfident" (above or equal to mean accuracy but below-mean confidence level). We found that a substantial portion of our sample was poorly calibrated. Specifically, 110 participants (22.4%) were overconfident and 135 participants (27.6%) were underconfident. Together, these findings reveal that nearly half of the participants exhibited a mismatch between their perceived and actual ability to identify AI-manipulated videos.

4.2.2 Factors Associated with Detection Accuracy. To understand what factors contributed to participants' detection accuracy, we conducted a series of analyses examining the influence of demographics, media habits, video characteristics, and self-reported trust and skepticism on videos encountered.

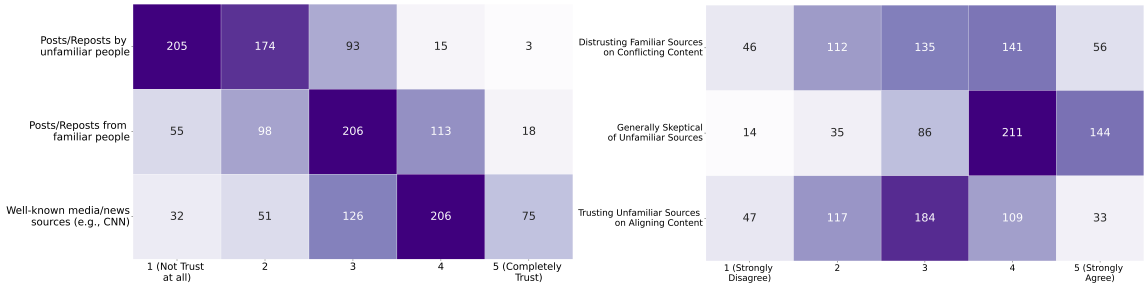
Our analyses of demographic and media habits revealed no statistically significant differences in accuracy based on age group (t-test: ≤ 35 vs. >35, $p = 0.94$; ANOVA: across five age-subgroups, $p = 0.95$), disability status (t-test: yes vs. no, $p = 0.90$), or daily social media exposure (t-test: ≤ 2 hours vs. >2 hours, $p = 0.44$).

Several other factors were significantly associated with performance in our sample. Student's t-tests showed that male participants ($M = 67.6\%, N = 238$) were more accurate than female participants ($M = 64.9\%, N = 249; p < .05$). Similarly, we found strong evidence for better detection accuracy among people without a college degree ($M = 68.1\%, N = 170$) vs people with a college degree or above ($M = 65.3\%, N = 320; p < .05$). Political affiliation and primary news source also correlated with accuracy. Participants identifying as Democrat and leaning-democrat ($M = 68.3\%, N = 215$) were more accurate than those identifying as Republican and leaning-republican ($M = 63.6\%, N = 182; p < .01$), and those who primarily consumed news from internet sources (e.g., social media, news websites & apps) ($M = 66.8\%, N = 412$) were more accurate than those who relied on traditional media like television or radio ($M = 63.2\%, N = 78; p < .05$).

Shifting from participant traits to video content, a Chi-square test revealed that participants were significantly more accurate when evaluating videos featuring famous individuals compared to those with unknown people ($p < .05$), which may suggest that participants' higher level of familiarity with the appearance and mannerisms of famous individuals help them in spotting manipulation in the videos. However, we did not see a significant impact from camera being stable vs. continuously moving in two similar videos ($p = 0.06$), variation in video quality (authentic pair: $p = 0.11$; manipulated pair: $p = 0.30$; combined: $p = 0.90$), and the inclusion of news organization logo (authentic pair: $p = 0.33$; manipulated pair, $p = 0.32$; combined: $p = 1$).

Another factor that plays a central role in the online world is the trust in content sources. On a 5-point scale, participants reported the highest trust in established media outlets and news applications ($M = 3.49, SD = 1.08$), lower trust in videos from people they follow or familiar with ($M = 2.88, SD = 1.01$), and the least trust in unfamiliar individuals ($M = 1.85, SD = 0.87$) (detail selection in Fig. 3(a)). This pattern suggests an unsurprising caution toward online content, especially from unfamiliar sources. To examine whether reported trust affected detection accuracy, we grouped ratings into low (1 & 2) and high (4 & 5) trust. Student's t-test results showed that participants with lower trust for familiar sources on social media achieved significantly higher accuracy in detecting manipulated media (66.3%, $N = 153$) than those with higher trust (62.0%, $N = 131; p < .01$).

And lastly, we considered confirmation bias [39]. On a 5-point scale, participants expressed strongest agreement with being generally skeptical of unfamiliar sources ($M = 3.89, SD = 1.00$), compared to trusting unfamiliar sources that aligned with beliefs ($M = 3.10, SD = 1.16$) or distrusting familiar sources when content conflicted with beliefs ($M = 2.93, SD = 1.05$) (detail selection in Fig. 3(b)). However, Student's t-tests on low versus high level of agreement with these statements showed no significant effect on detection accuracy.



(a) Trust level on content sources

(b) Agreement with statements on trusting video content based on source familiarity and belief alignment

Fig. 3. Heatmaps showing the distribution of two 5-point Likert Scale questions in the survey.

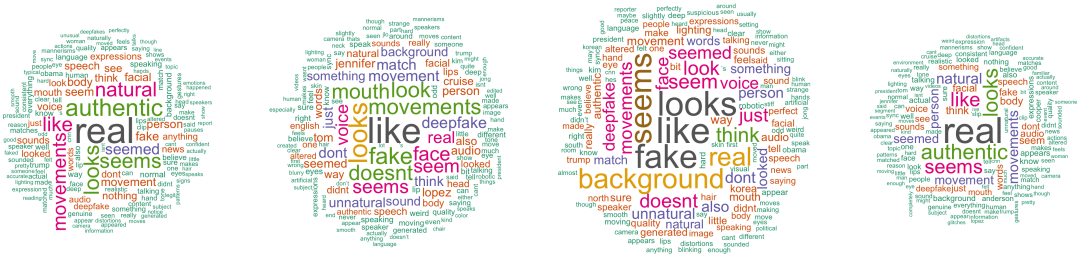
4.2.3 Relied Upon Cues and Evaluation Strategies. To understand the specific cues in a video that participants rely on when evaluating a video's authenticity, we asked them to rate the influence of various factors on their suspicion. Our findings indicate that participants consider a wide range of factors to be influential, with all eight presented factors receiving a mean score near 4 on a 5-point scale (detailed results shown in Fig. 4). However, participants rated some factors as more critical to their evaluation strategy than others. *Audio/video synchronization* was rated as the most critical factor ($M = 4.47, SD = 0.79$), followed closely by *facial expressions* ($M = 4.38, SD = 0.86$) and *body movements* ($M = 4.32, SD = 0.91$). In contrast, *audio artifacts* were rated as the least influential factor ($M = 3.77, SD = 1.04$).

Fig. 4. Heatmap showing the distribution of participants' ratings of factors influencing their suspicion that a video might not be authentic.

When asked which artifact they typically notice first, a significant portion of participants (44.9%, $N = 220$) selected *facial expression issues.* This aligns with the technical nature of many common AI-driven video manipulation methods, such as face-swapping or reenactment manipulations [36], which focus on altering the facial region and can introduce

unnatural-looking expressions. Following this, *audio synchronization issues* were the next most frequently chosen first-noticed artifact (28.4%, $N = 139$). The prominence of this multi-modal cue is notable when compared to single-modality flaws like *unusual visual artifacts* (17.1%, $N = 84$) or *unusual audio artifacts* (4.7%, $N = 23$). Very few participants (4.9%, $N = 24$) reported that *low video or audio quality* was the first thing they would notice. As generative AI technology continues to advance, the quality of video and audio output is improving, and the general reluctance to heavily rely on quality could potentially be a major resilience factor.

To contextualize these self-reported strategies, we analyzed the open-ended reasoning participants provided for their judgments on each video. The word clouds generated from these responses (see Fig. 5) reveal that *movement* has been a common term in all scenarios, along with some mentions of *expression*, supporting our quantitative finding that participants focus heavily on human mannerisms. Participants' reasoning also highlighted a focus on specific locations, with terms such as *mouth*, *face*, *body*, and *background* appearing frequently. Similarly, the presence of words like *audio*, *speech*, *voice*, and *sounds* in their justifications reinforces the importance of the audio track in their evaluations and could connect with their high rating of audio/video synchronization as a critical detection cue.



(a) Correctly identified authentic videos (b) Correctly identified manipulated videos (c) Wrongly identified authentic videos (d) Wrongly identified manipulated videos

Fig. 5. Word clouds capturing the responses participants left justifying their judgment on the seen videos.

4.3 RQ3: Verification Practices and Tool Awareness

4.3.1 Common Actions Taken in Response to Suspected AI-manipulated Videos. Beyond the ability to judge a video's authenticity, our study also investigated what actions individuals would take when encountering a suspected AI-manipulated video on social media through a multiple choice question. We present the size of each individual choice and common combinations (with 10+ selections) in Fig. 6.

Our findings reveal a split between passive disdain and proactive verification. A significant portion of participants (27.1%, $N = 133$) indicated a solely passive approach, stating they would simply ignore the content. Among those who would take proactive steps, the most common solely-used strategy is to search the internet for other sources (24.9%, $N = 122$), followed by checking user comments and responses (7.3%, $N = 36$). The combination of these two strategies was also a frequent choice, selected by 86 participants. In stark contrast, using a dedicated video validation website or detection tool was the least common action, chosen by only four participants as their only strategy and by 13 as part of any combination. Passivity being the most common response might raise questions about when ignoring a suspected video is appropriate. Given the modest detection accuracy observed among participants, it does not seem to

RQ2 Key Insights

Modest accuracy with poorly calibrated confidence. Participants' overall detection accuracy was better than chance but not reliably high, sharing a similar finding with Diel et al. [17]. However, regarding self-assessment, our findings diverge slightly from prior work that typically observes a general trend of overconfidence [26, 58]. In our sample, about half (i.e., 22.4% overconfident and 27.6% underconfident) of the participants demonstrated poorly calibrated confidence levels in their detection ability, suggesting a need to explore ways to help people better align their perceived ability with their actual performance.

No "Secret Formula" for Detection Ability. People's ability to detect AI-manipulated videos is a complex problem, and vulnerability appears to be relatively universal. While prior work has examined demographic predictors of detection ability, identified influence could be mixed [59]. In our specific context, commonly assumed influential factors, such as age and social media exposure, did not show significant effect on detection accuracy among participants. Moreover, additional years of traditional education does not seem to improve ability within this context–on the contrary, people with a college degree demonstrated lower accuracy than those without in our study.

Evaluation focuses on human-centric cues. Consistent with prior research indicating that observers prioritize the facial region when assessing authenticity [64, 69], we found our participants were guided more by human-centric, semantic cues like unnatural facial expressions and body movements than technical artifacts like poor video quality or visual/audio artifacts. This finding shows that people primarily search for violations of "humanness," using their intuitive sense of authentic behavior as the main evidence.

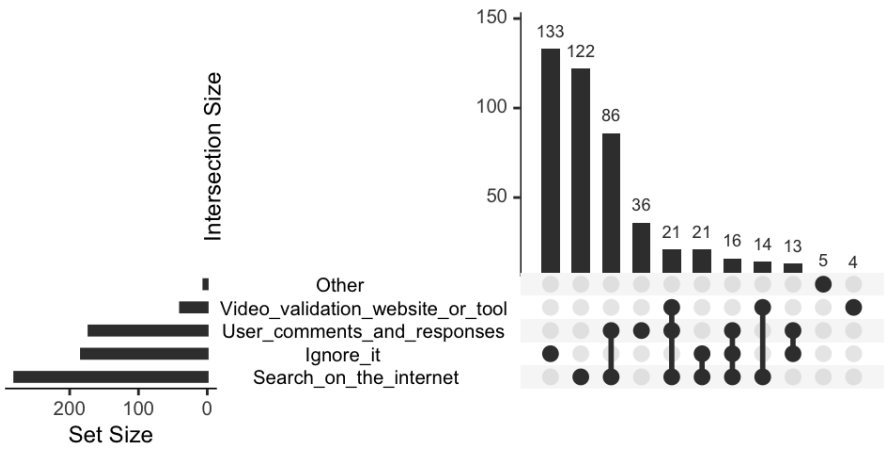


Fig. 6. An UpSet plot showing the size of each individual actions participants would take facing suspected AI-manipulated videos and common combination of actions.

be the ideal response here. However, when or why users default to this approach, or whether it is reasonable to expect them to perform active verification are interesting questions that could be examined in future research.

To understand what distinguishes participants who ignore suspected content from those who investigate it, we modeled the decision using a binary logistic regression. The dependent variable categorized participants into two groups: solely ignoring the content (passive) versus taking at least one verification action (proactive). The model included the same set of demographic and media habit predictors established in Sec. 4.1.2, with the addition of perceived prevalence of AI-media. We found two significant predictors of proactive verification behavior. First, daily social media usage

was a strong predictor ($Est. = 0.62, p = 0.004, OR = 1.86$). Participants who spend more than two hours daily on social media had nearly double the odds of taking proactive verification steps compared to lighter users. Second, the perceived use of the technology played a significant role ($Est. = 0.50, p = 0.03, OR = 1.65$). Participants who viewed the primary application of AI-manipulated videos as "malicious" had 1.65 times higher odds of verifying content compared to those who viewed it primarily as "entertainment."

Finally, the open-ended responses in the "Other" category provided further insight into alternative strategies. Out of 14 participants, who specified actions that were not listed, 10 participants suggested they would report the content to the platform. This highlights a desire for platform-level intervention, although this desire was seen to be paired with skepticism, as one participant noted, *"report it...not that social media companies ever do anything with reports..."*. Another participant mentioned they *"may potentially block the user,"* a strategy focused on personal content curation rather than broader verification.

4.3.2 Public Awareness and Usage of Detection Tools. To assess the public's familiarity with technological solutions, we asked participants about their awareness of specialized detection tools for AI-manipulated videos. The results reveal a significant awareness gap: an overwhelming majority of participants (91.5%, $N = 458$) were not aware of any tools designed to analyze and verify AI-manipulated videos. Only 8.5% ($N = 42$) indicated that they knew of such tools. Among this small subset, several specialized tools were named, including Deepware (mentioned by 10 participants) [14], Microsoft Video Authenticator (7) [8], Sensity AI (4) [3], and Reality Defender (3) [15]. Two participants also listed Snopes [55], suggesting the usage of established fact-checking websites over specialized technical tools for verification.

Furthermore, our findings show that awareness does not directly translate to usage. Of the 42 participants who knew of detection tools, less than half (N=17) reported that they would actually use one to check a suspicious video's authenticity. This indicates that barriers beyond simple awareness, such as usability or accessibility, might be inhibiting adoption, which requires further exploration.

Compounding this issue is a sign of confusion about the nature of available AI technologies. Four participants mistakenly identified generative AI tools (e.g. Sora [43], Grok [71]) as detection tools. While this number is small, it points to a broader challenge where the constant stream of "buzz words" in public discourse may make it difficult for individuals to distinguish between tools that create AI media and those designed to detect it.

RQ3 Key Insights

Engagement is linked to exposure and threat perception. While a significant portion of participants default to ignoring suspected content, our regression analysis reveals that heavy social media usage and viewing AI-manipulated media as a malicious threat increase the odds of choosing proactive verification.

Participants look for platform-level interventions. Reporting content to the platform was a notable action brought up by participants. Some platforms experiment with AI-generated content labels [21], this finding points to a design opportunity to move from reporting to exploring more collaborative systems that leverage user input to counter manipulated media [46].

Detection tools are not widely known or utilized. Despite the very active research on and rapid development of detection technologies, awareness and usage of resulting tools remain low among the participants. Future research is needed to understand the specific barriers to adoption, such as concerns about tool reliability, accessibility and usability, in order to bridge this gap.

5 Limitations

Our experimental design and methodology had limitations that should be considered when interpreting the results.

Methodological Constraints. Due to the online nature of the survey, we had limited control over the viewing environment and potential distractions. While this may reflect real-life conditions, we cannot describe or quantify the actual environments people took the survey in. Despite efforts to minimize them, our results subject to potential biases [45], especially the self-selection, self-reporting and response biases, such as social-desirability and observer effect. Moreover, participants were explicitly tasked with evaluating authenticity, which likely primed them to look for manipulation and may have inflated detection accuracy relative to a naturalistic setting.

Generalizability and Scope. We also caution against over-interpreting the results beyond the specific context of this study. First, our sample was limited to U.S. adults, and cultural differences regarding the perception of the social media landscape and technology usage may limit the generalizability of the findings. Second, although using in-the-wild stimuli improved ecological validity, the 23 AI-manipulated videos cannot capture the full diversity of content on social media. Finally, given the rapidly changing nature of AI tools for manipulating and generating videos, the specific detection rates reported here should be viewed as a snapshot in time, even though we expect the underlying behavioral patterns related to suspicion and verification to remain relevant.

6 Future Work

Our findings point to two promising directions for future research to further understand and mitigate the challenges posed by AI-manipulated media.

First, while our quantitative analysis reveals patterns, such as the gender disparity in perceived prevalence and the notably low awareness of detection tools, survey data alone cannot fully capture the nuance of *why* these patterns emerge. Future qualitative approaches, such as semi-structured interviews or focus groups, could help dissect the underlying thought processes that drive these behaviors. Such an investigation could reveal, for example, whether the limited tool usage comes from a lack of trust in the technology, usability hurdles, or simply because of the disruption to people's browsing habits.

Second, expanding this methodology to cross-cultural contexts would help address the limitations of a U.S.-based sample. Since media ecosystems vary globally, comparative studies across nations with differing social media landscapes and levels of digital literacy are important to understand how these environmental factors shape public perception, detection ability, and verification behaviors.

7 Conclusion

This paper presents a demographically balanced study of 490 U.S. adults to characterize the public's relationship with AI-manipulated videos using in-the-wild video stimuli. Our findings reveal a complex landscape: while perceptions of prevalence are widespread yet diverse, the public's actual ability to identify manipulated content remains limited, marked by a striking mismatch between confidence and performance. Evaluation strategies rely predominantly on intuitive, human-centric cues rather than technical artifacts, and awareness and adoption of specialized detection tools remains low despite substantial recent research and investment. As the line between authenticity and manipulated reality continues to blur, fostering a more resilient public will require not only stronger technical solutions but also sociotechnical interventions that educate and empower users in their daily digital interactions.

References

[1] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. 2020. Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2814–2822. doi:10.1109/CVPRW50498.2020.00338

[2] Saifuddin Ahmed. 2023. Navigating the maze: Deepfakes, cognitive ability, and social media news skepticism. *New Media & Society* 25, 5 (2023), 1108–1129.

[3] Sensity AI. [n. d.]. Deepfakes Detection. https://sensity.ai/deepfake-detection. [Accessed 09-06-2025].

[4] Yvonne Apolo and Katina Michael. 2024. Beyond a reasonable doubt? Audiovisual evidence, AI manipulation, deepfakes, and the law. *IEEE Transactions on Technology and Society* 5, 2 (2024), 156–168.

[5] Natalie Grace Brigham, Miranda Wei, Tadayoshi Kohno, and Elissa M. Redmiles. 2024. "Violation of my body": perceptions of AI-generated non-consensual (intimate) imagery. In *Proceedings of the Twentieth USENIX Conference on Usable Privacy and Security* (Philadelphia, PA, USA) *(SOUPS '24)*. USENIX Association, USA, Article 20, 20 pages.

[6] Catherine Francis Brooks. 2021. Popular discourse around deepfakes and the interdisciplinary challenge of fake video distribution. *Cyberpsychology, Behavior, and Social Networking* 24, 3 (2021), 159–163.

[7] US Census Bureau. [n. d.]. Data — census.gov. https://www.census.gov/data.html. [Accessed 09-09-2025].

[8] Tom Burt. 2020. New steps to combat disinformation. [Accessed 09-06-2025].

[9] Akash Chintha, Aishwarya Rao, Saniat Sohrawardi, Kartavya Bhatt, Matthew Wright, and Raymond Ptucha. 2020. Leveraging Edges and Optical Flow on Faces for Deepfake Detection. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*. 1–10. doi:10.1109/IJCB48548.2020.9304936

[10] Beomsang Cho, Binh M. Le, Jiwon Kim, Simon Woo, Shahroz Tariq, Alsharif Abuadbba, and Kristen Moore. 2023. Towards Understanding of Deepfake Videos in the Wild. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (Birmingham, United Kingdom) *(CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 4530–4537. doi:10.1145/3583780.3614729

[11] Alex Clark. 2025. Deepfake videos impersonating real doctors push false medical advice and treatments — cbsnews.com. https://www.cbsnews.com/news/deepfake-videos-impersonating-real-doctors-push-false-medical-advice-treatments/. [Accessed 09-06-2025].

[12] Justin D Cochran and Stuart A Napshin. 2021. Deepfakes: awareness, concerns, and platform accountability. *Cyberpsychology, Behavior, and Social Networking* 24, 3 (2021), 164–172.

[13] DeepTrace. 2019. The State of Deepfakes: Landscape, Threats, and Impact. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf. [Accessed 09-06-2025].

[14] deepware.ai. [n. d.]. Deepware | Scan & Detect Deepfake videos. https://deepware.ai. [Accessed 09-04-2025].

[15] Reality Defender. [n. d.]. Reality Defender. https://realitydefender.com. [Accessed 09-06-2025].

[16] Nicholas Diakopoulos and Deborah Johnson. 2021. Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New media & society* 23, 7 (2021), 2072–2098.

[17] Alexander Diel, Tania Lalgi, Isabel Carolin Schröter, Karl F. MacDorman, Martin Teufel, and Alexander Bäuerle. 2024. Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports* 16 (2024), 100538. doi:10.1016/j.chbr.2024.100538

[18] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397* (2020).

[19] Sindhu Kiranmai Ernala, Moira Burke, Alex Leavitt, and Nicole B. Ellison. 2020. How Well Do People Report Time Spent on Facebook? An Evaluation of Established Survey Questions with Recommendations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376435

[20] Dilrukshi Gamage, Piyush Ghasiya, Vamshi Bonagiri, Mark E. Whiting, and Kazutoshi Sasahara. 2022. Are Deepfakes Concerning? Analyzing Conversations of Deepfakes on Reddit and Exploring Societal Implications. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 103, 19 pages. doi:10.1145/3491102.3517446

[21] Dilrukshi Gamage, Dilki Sewwandi, Min Zhang, and Arosha K Bandara. 2025. Labeling Synthetic Content: User Perceptions of Label Designs for AI-Generated Content on Social Media. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 814, 29 pages. doi:10.1145/3706598.3713171

[22] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. 2022. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences* 119, 1 (2022), e2110013119.

[23] Matthew Groh, Aruna Sankaranarayanan, Nikhil Singh, Dong Young Kim, Andrew Lippman, and Rosalind Picard. 2024. Human detection of political speech deepfakes across transcripts, audio, and video. *Nature communications* 15, 1 (2024), 7629.

[24] Muhammad Riyyan Khan, Shahzeb Naeem, Usman Tariq, Abhinav Dhall, Malik Nasir Afzal Khan, Fares Al Shargie, and Hasan Al-Nashash. 2023. Exploring Neurophysiological Responses to Cross-Cultural Deepfake Videos. In *Companion Publication of the 25th International Conference on Multimodal Interaction* (Paris, France) *(ICMI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 41–45. doi:10.1145/3610661.3617148

[25] Jan Kietzmann, Linda W Lee, Ian P McCarthy, and Tim C Kietzmann. 2020. Deepfakes: Trick or treat? *Business horizons* 63, 2 (2020), 135–146.

[26] Nils C Köbis, Barbora Doležalová, and Ivan Soraperra. 2021. Fooled twice: People cannot detect deepfakes but think they can. *Iscience* 24, 11 (2021).

[27] Pavel Korshunov and Sébastien Marcel. 2020. Deepfake detection: humans vs. machines. *arXiv preprint arXiv:2009.03155* (2020).

[28] Mateusz Łabuz and Christopher Nehring. 2024. On the way to deep fake democracy? Deep fakes in election campaigns in 2023. *European Political Science* 23, 4 (2024), 454–473.

[29] Andrew Lewis, Patrick Vu, Raymond M Duch, and Areeq Chowdhury. 2023. Deepfake detection with and without content warnings. *Royal Society Open Science* 10, 11 (2023), 231214.

[30] Xialing Lin, Patric R. Spence, and Kenneth A. Lachlan. 2016. Social media and credibility indicators: The effect of influence cues. *Computers in Human Behavior* 63 (2016), 264–271. doi:10.1016/j.chb.2016.05.002

[31] Nikola Marangunić and Andrina Granić. 2015. Technology acceptance model: a literature review from 1986 to 2013. *Universal access in the information society* 14, 1 (2015), 81–95.

[32] Marie-Helen Maras and Alex Alexandrou. 2019. Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. *The international journal of evidence & proof* 23, 3 (2019), 255–262.

[33] Vineet Mehta, Parul Gupta, Ramanathan Subramanian, and Abhinav Dhall. 2021. FakeBuster: A DeepFakes Detection Tool for Video Conferencing Scenarios. In *Companion Proceedings of the 26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21 Companion)*. Association for Computing Machinery, New York, NY, USA, 61–63. doi:10.1145/3397482.3450726

[34] Bradley D Menz, Natansh D Modi, Michael J Sorich, and Ashley M Hopkins. 2024. Health disinformation use case highlighting the urgent need for artificial intelligence vigilance: weapons of mass disinformation. *JAMA internal medicine* 184, 1 (2024), 92–96.

[35] Jaron Mink, Licheng Luo, Natã M. Barbosa, Olivia Figueira, Yang Wang, and Gang Wang. 2022. DeepPhish: Understanding User Trust Towards Artificially Generated Profiles in Online Social Networks. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Boston, MA, 1669–1686. https://www.usenix.org/conference/usenixsecurity22/presentation/mink

[36] Yisroel Mirsky and Wenke Lee. 2021. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.* 54, 1, Article 7 (Jan. 2021), 41 pages. doi:10.1145/3425780

[37] Rafał Muda, Gordon Pennycook, Damian Hamerski, and Michał Białek. 2023. People are worse at detecting fake news in their foreign language. *Journal of Experimental Psychology: Applied* 29, 4 (2023), 712.

[38] Fabian Muhly, Emanuele Chizzonic, and Philipp Leo. 2025. AI-deepfake scams and the importance of a holistic communication security strategy. *International Cybersecurity Law Review* (2025), 1–9.

[39] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175–220.

[40] University of Buffalo. [n. d.]. DeepFake-o-Meter. [Accessed 09-04-2025].

[41] U.S. Department of Labor. [n. d.]. Minimum Wage. https://www.dol.gov/agencies/whd/minimum-wage. Accessed 12-04-25.

[42] Ofcom. [n. d.]. Ofcom Media Plurality Quantitative Questionnaire. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/multi-sector/media-plurality/quantitative-research-questionnaire?v=328783. Accessed 11-26-2025.

[43] OpenAI. [n. d.]. Sora. https://openai.com/sora. [Accessed 09-06-2025].

[44] Optic. [n. d.]. AI or Not. https://www.aiornot.com. [Accessed 09-04-2025].

[45] Ganna Pogrebna, Karen Renaud, and Marina Kovaleva. 2025. *Big Bad Bias Book: A Field Guide to Over 200 Cognitive Biases That Shape How We Think, Decide, and Behave.*

[46] Adam Presser. [n. d.]. Rolling out TikTok Footnotes in the US. https://newsroom.tiktok.com/en-us/rolling-out-tiktok-footnotes-in-the-us. [Accessed 09-09-2025].

[47] Prolific. [n. d.]. Prolific | Easily collect high-quality data from real people. https://www.prolific.com.

[48] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 86–103.

[49] Qualtrics. [n. d.]. Qualtrics XM - Experience Management Software. https://www.qualtrics.com. [Accessed 09-09-2025].

[50] Surfshark Research. 2025. Deepfake fraud caused financial losses nearing $900 million — surfshark.com. https://surfshark.com/research/chart/deepfake-fraud-losses. [Accessed 09-06-2025].

[51] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[52] Margie Ruffin, Haeseung Seo, Aiping Xiong, and Gang Wang. 2024. Does It Matter Who Said It? Exploring the Impact of Deepfake-Enabled Profiles on User Perception towards Disinformation. *Proceedings of the International AAAI Conference on Web and Social Media* 18, 1 (May 2024), 1328–1341. doi:10.1609/icwsm.v18i1.31392

[53] Sensity. 2024. The State of Deepfakes 2024. https://sensity.ai/reports/. [Accessed 09-06-2025].

[54] Farhana Shahid, Srujana Kamath, Annie Sidotam, Vivian Jiang, Alexa Batino, and Aditya Vashistha. 2022. " It Matches My Worldview": Examining Perceptions and Attitudes Around Fake Videos. In *proceedings of the 2022 CHI conference on human factors in computing systems*. 1–15.

[55] Snopes. [n. d.]. Snopes. https://www.snopes.com. [Accessed 09-06-2025].

[56] Saniat Javid Sohrawardi, Akash Chintha, Bao Thai, Sovantharith Seng, Andrea Hickerson, Raymond Ptucha, and Matthew Wright. 2020. DeFaking Deepfakes: Understanding Journalists' Needs for Deepfake Detection. In *Computation + Journalism Symposium*.

[57] Saniat Javid Sohrawardi, Y. Kelly Wu, Andrea Hickerson, and Matthew Wright. 2024. Dungeons & Deepfakes: Using scenario-based role-play to study journalists' behavior towards using AI-based verification tools for video content. In *Proceedings of the 2024 CHI Conference on Human*

Factors in Computing Systems (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 776, 17 pages. doi:10.1145/3613904.3641973

[58] Klaire Somoray and Dan J. Miller. 2023. Providing detection strategies to improve human detection of deepfakes: An experimental study. *Computers in Human Behavior* 149 (2023), 107917. doi:10.1016/j.chb.2023.107917

[59] Klaire Somoray, Dan J Miller, and Mary Holmes. 2025. Human Performance in Deepfake Detection: A Systematic Review. *Human Behavior and Emerging Technologies* 2025, 1 (2025), 1833228.

[60] Daniel Story and Ryan Jenkins. 2023. Deepfake pornography and the ethics of non-veridical representations. *Philosophy & Technology* 36, 3 (2023), 56.

[61] The Sumsuber. 2023. The Top KYC Trends Coming in 2024. https://sumsub.com/blog/top-kyc-trends/. [Accessed 09-06-2025].

[62] The Sumsuber. 2024. 2024 Identity Theft & Fraud Statistics. https://sumsub.com/fraud-report-2024/. [Accessed 09-06-2025].

[63] Stefan Sütterlin, Torvald F Ask, Sophia Mägerle, Sandra Glöckler, Leandra Wolf, Julian Schray, Alava Chandi, Teodora Bursac, Ali Khodabakhsh, Benjamin J Knox, et al. 2023. Individual deep fake recognition skills are affected by viewer's political orientation, agreement with content and device used. In *International Conference on Human-Computer Interaction*. Springer, 269–284.

[64] Rashid Tahir, Brishna Batool, Hira Jamshed, Mahnoor Jameel, Mubashir Anwar, Faizan Ahmed, Muhammad Adeel Zaffar, and Muhammad Fareed Zaffar. 2021. Seeing is Believing: Exploring Perceptual Differences in DeepFake Videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 174, 16 pages. doi:10.1145/3411764.3445699

[65] Paulina Trifonova and Sukrit Venkatagiri. 2024. Misinformation, Fraud, and Stereotyping: Towards a Typology of Harm Caused by Deepfakes. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing* (San Jose, Costa Rica) *(CSCW Companion '24)*. Association for Computing Machinery, New York, NY, USA, 533–538. doi:10.1145/3678884.3685938

[66] Rebecca Umbach, Nicola Henry, Gemma Faye Beard, and Colleen M. Berryessa. 2024. Non-Consensual Synthetic Intimate Imagery: Prevalence, Attitudes, and Knowledge in 10 Countries. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 779, 20 pages. doi:10.1145/3613904.3642382

[67] Georgetown University. [n. d.]. TrueMedia.org. https://www.truemedia.org. [Accessed 09-07-2025].

[68] Cristian Vaccari and Andrew Chadwick. 2020. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society* 6, 1 (2020), 2056305120903408. arXiv:https://doi.org/10.1177/2056305120903408 doi:10.1177/2056305120903408

[69] Leslie Wöhler, Martin Zembaty, Susana Castillo, and Marcus Magnor. 2021. Towards Understanding Perceptual Differences between Genuine and Face-Swapped Videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 240, 13 pages. doi:10.1145/3411764.3445627

[70] Y. Kelly Wu, Saniat Javid Sohrawardi, Candice R. Gerstner, and Matthew Wright. 2025. Understanding and Empowering Intelligence Analysts: User-Centered Design for Deepfake Detection Tools. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 870, 26 pages. doi:10.1145/3706598.3713711

[71] xAI. [n. d.]. Grok. https://grok.com. [Accessed 09-6-2025].

[72] Aya Yadlin-Segal and Yael Oppenheim. 2021. Whose dystopia is it anyway? Deepfakes and social media regulation. *Convergence* 27, 1 (2021), 36–51.

[73] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. 2023. UCF: Uncovering Common Features for Generalizable Deepfake Detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 22355–22366. doi:10.1109/ICCV51070.2023.02048

[74] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. 2023. DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 4534–4565. https://proceedings.neurips.cc/paper_files/paper/2023/file/0e735e4b4f07de483cbe250130992726-Paper-Datasets_and_Benchmarks.pdf

[75] Zewei Zong. [n. d.]. Eliminate order bias to improve your survey responses. https://www.surveymonkey.com/curiosity/eliminate-order-bias-to-improve-your-survey-responses. [Accessed 09-10-2025].

A Survey - Part 1

Please indicate your age range:

- 19 - 24
- 25 - 34
- 35 - 44
- 45 - 54
- 55 or Above

What is the highest degree you have completed?

- Less than high school diploma
- High school diploma or equivalent
- Some college (no degree)
- Associate Degree
- Bachelor's Degree
- Master's Degree
- Doctoral Degree

Which of the following best describes your gender?

- Male
- Female
- Non-binary / Other
- Prefer not to say

Do you have any disabilities?

- Yes
  - What type of disability or impairment you have?
    - Visual
    - Hearing
    - Cognitive
    - Other –> {text entry}
    - Prefer not to say
- No
- Prefer not to say

Which race best describes you?

- American Indian or Alaska Native
- Asian
- Black or African American
- Hispanic
- Native Hawaiian or Other Pacific Islander
- White / Caucasian
- Other –> {text entry}
- Prefer not to say

On an average how many hours per day you spent on social media platforms such as Facebook, YouTube, TikTok, Truth Social, Instagram, Online Forums, etc.?

- less than 30min
- 30min to 1hr
- 1-2 hours
- 2-3 hours
- more than 3 hours

In digital world, where do you get your news from? Please rank the choices below from most common source for you to the least common by sliding the choices below (where top (rank 1) choice represents where you get most of your news from on a typical day)

- Social Media (X/Twitter, Facebook, Instagram, Truth Social, Online Forums, Youtube etc.)
- Aggregated News Apps/Sites (Google News, Apple News, Flipboard, etc)
- Website(s), Apps or podcasts of mainstream newspaper or news networks (CNN, Fox, NY Times, Financial Times, etc.)
- Television or Radio

Deepfakes are defined as multimedia content that are generated by Artificial Intelligence, or that are altered to misrepresent someone or their environment as doing or saying something that was not actually done or said as represented in the content. What percentage of multimedia content (videos, pictures, audio) on social media you think consists of deepfakes?

- Less than 5
- 6-10
- 11-20
- 21-30
- 31-40
- 41-50
- More than 50

There are many social networks, please choose the answer that starts with letter F. This question is very simple to check your attention.

- Facebook
- Instagram
- X/Twitter
- TikTok

What do you think is the most common use of deepfakes today?

- Non-consensual pornography[4]
- Cyber bullying or harassment
- Entertainment
- Political misinformation or election interference
- Scientific or medical misinformation

[4]We only use "pornography" rather than NCII in the survey because it is more widely recognized by the general public.

- Blackmailing, financial fraud or scams
- Other –> {text entry}

What best describes your political view or leaning in the U.S. Politics?

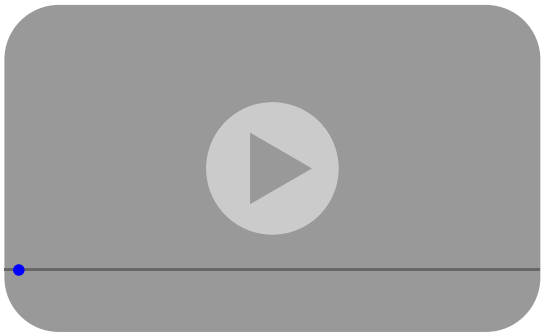
- Republican / Leaning Republican
- Democrat / Leaning Democrat
- Other –> {text entry}
- Prefer not to say

B Survey - Part 2

(Each participant filled this part of the Survey 13 times, once for each video they were presented with)

Please use the reference definitions below when you answering the question below:
Authentic Video: A video showing actual people, events, or actions as it happened.
Deepfake Video: A video that is generated by Artificial Intelligence or it is altered to misrepresent someone or their environment as doing or saying something that was not actually done or said as represented in the video.

Do you think the video you just watched is a deepfake or authentic?



- Deepfake
- Authentic

How confident do you feel about your answer? Rate your estimated Confidence level on the scale from 1 (Not Confident at all) to 5 (Extremely Confident) below:

- {5-point likert scale}

Why did you make this decision about the video? Please explain your reasoning in 1–2 sentences.

- {text entry}

C Survey - Part 3

To what extent do the following factors in a video influence your suspicion that it might be a deepfake? Rate each on the scale from 1 (Not at all influential) to 5 (Extremely influential)

- Overall visual quality
- Overall audio quality
- Audio/Video synchronization
- Audio artifacts
- Visual artifacts
- Facial expressions
- Body movements
- Background or environment

Which of the following you typically notice first on a deepfake video? (Select one)

- Audio synchronization issues
- Low video or audio quality
- Unusual audio artifacts
- Facial expression issues
- Unusual visual artifacts

How does the platform or the person sharing it impacts your trust on a video? Rate your estimated trust level on the scale from 1 (do not trust at all) to 5 (completely trust) for each case below:

- Well-known media/news websites or mobile apps (e.g., CNN, Fox, NBC, etc.)
- Posts or reposts on Social media by people you **follow or familiar** with
- Posts or reposts on Social media by people you are **NOT familiar** with

Please indicate the extent to which you agree or disagree with the following statements. Use the scale from 1 (Strongly Disagree) to 5 (Strongly Agree)

- If a video conflicts my existing knowledge, views, or beliefs, I'm less likely to trust it—even if it's from a usually **reliable and familiar** source.
- I'm generally skeptical of videos from **unfamiliar** sources, no matter what the content is.
- If a video aligns with my existing knowledge, views, or beliefs, I'm more inclined to trust it—even if the source is **unfamiliar**
- I swim across the pacific ocean to get to work **everyday**

What would you typically do if you saw a video on a topic you care about on social media but suspect that it is a deepfake? (multiple choice)

- I would just ignore it
- I would search on the internet for other sources and try to fact-check it
- I would check the user comments and responses to see what others think
- I would use a video validation website or a tool to check whether it is a deepfake
- Other (please specify) –> {text entry}

Do you know of any tools that you can use to analyze video to check whether it is deepfake? If so, please also name them below

- Yes –> {text entry}
- No

If you suspect a video is deepfake, do you typically use any video analysis tool to check its authenticity? If yes, please name the one you use the most?

- Yes –> {text entry}
- No
- I don't know of any such tool