Laypeople's Knowledge and Concerns Regarding Deepfakes and Other Manipulated Media

Wentao Guo Diana Chou Kamala Varma
University of Maryland, College Park

Abstract

Deepfakes—powerful tools for manipulating videos realistically—have been the subject of much concern, but this has largely focused on celebrities, politicians, and other public figures. We conducted semi-structured interviews with six U.S. adults to understand laypeople's knowledge and attitudes about manipulated videos and photos. We find that our participants considered themselves unlikely to be depicted without consent in manipulated videos or photos, but most would be highly concerned if they were. They generally expressed a lack of technical understanding about how deepfakes work, but some were familiar with existing types of deepfakes. However, participants also expressed misconceptions about deepfakes that could have negative implications for effectively protecting themselves. We conclude with recommendations to mitigate potential harms caused by manipulated media.

1 Introduction

A variety of recent technologies, grouped under the umbrella term *deepfakes*, have made it easier and more accessible for non-experts to create realistic synthetic videos that convincingly depict other people. Public attention from media, lawmakers, and researchers have focused on harms caused by deepfakes, both potential and actively occurring. Commonly discussed potential harms include disinformation targeting social media users, heads of state, and vulnerable populations such as children [4]; actively occurring harms that have received much attention primarily comprise fake nudes and pornography, typically depicting female celebrities and public figures [6].

However, little investigation, academic or otherwise, has engaged with laypeople: people without widespread public recognition who may be targeted primarily as co-workers, acquaintances, or simply strangers on the Internet. At the same time, there is some evidence that deepfake tools have been spreading on the Internet [7] and are being used by large numbers of people to create fake nudes and pornography of laypeople, not just public figures [2]. Yet, we know little about laypeople's knowledge, experiences, and concerns about deepfakes that might seek to depict them.

We conducted semi-structured interviews with six U.S. adults, seeking to answer the following research questions:

1. What do people currently know about deepfakes and other manipulated media?
2. What are people's current concerns regarding potential harms resulting from non-consensual manipulated media?
3. What are people's reactions, attitudes, and concerns when prompted with information about manipulated media or descriptions of specific hypothetical scenarios?

This is the first part of a larger research project; while our current work asks participants to think through potential harms of deepfakes, we also intend to study the experiences of laypeople who have actually been depicted in deepfakes in harmful ways without their consent. We believe that filling in these gaps will help inform more effective policies, educational interventions, technological designs, and other solutions aimed at mitigating harm from deepfakes.

2 Related Work

To the best of our knowledge, this is the first work that focuses specifically on the current knowledge and concerns that laypeople hold regarding manipulated media and their harms. Prior work that relates to or motivates this topic can be divided into three main categories: empirical measurements of image-based sexual abuse, policies surrounding deepfakes, and the increasing accessibility of deepfake software.

2.1 Image-Based Sexual Abuse

Recent work has explored the experiences of victims of image-based sexual abuse (IBSA), defined as the "non-consensual creation and/or distribution of private sexual images," including deepfakes [15]. Large-scale surveys of Australian [12] and U.S. [17] adults indicate that IBSA is a prevalent issue that can be difficult for individuals to bring under control. Victims rarely seek help and may have significantly worse mental and physical health. Both surveys identified certain marginalized populations that were victimized disproportionately. Through our work, we intend to better inform solutions that will mitigate these various harms.

Bates [3] conducted in-depth interviews with IBSA survivors, finding that they reported a loss of trust, self-esteem, confidence, and sense of control, as well as diagnoses such as PTSD, anxiety, and depression. We hypothesize that survivors of non-consensual manipulated media may also share some of these negative mental health effects. Through our work, we hope to better understand how survivors' experiences may also differ due to the "fake" nature of manipulated media, in order to inform solutions that meet different survivors' preferences and needs.

In addition to being "fake," manipulated media differs from other types of IBSA in that the underlying technology is likely to be confusing or unfamiliar to survivors. Freed et al. [10] worked with survivors of intimate partner violence, who are often subject to technically confusing or unfamiliar spyware. They found that survivors tended to express concerns about technological surveillance vaguely, and clinical assessments with trained technologists were successful in uncovering security vulnerabilities and in engaging survivors. In our study, we also investigate how people communicate technical concerns about manipulated media. We hope that our findings may complement existing work on how to treat IBSA in clinical settings, which makes recommendations about inclusive and supportive language to empower survivors, but does not significantly consider the difficulty of communicating crucial technical details about manipulated media in particular [9].

2.2 Policies Surrounding Deepfakes

Previous work has explored the effects of deepfakes in varying contexts. De Ruiter examines the ethical implications of deepfake technology, arguing that it is not inherently morally wrong, even though the technology is susceptible to malicious activities and other actions that violate fundamental norms [8]. She argues that the technology can be used to reinforce people's autonomy, create entertainment, and allow for fictional curiosity. However, this is presented from an optimistic viewpoint, and we need to consider the potential negative ethical implications that arise from misuse of the technology. Burkell and Gosse emphasize the social and cultural context of deepfakes, looking at the harm inflicted on individuals targeted by non-consensual fake pornography in the era of the Internet, avatars, and digital photography [5]. Looking at the varying contexts in which individuals are subject to harm from deepfakes calls for better understanding of laypeople's perspectives.

Research has also been done in an effort to suggest some form of policy change or solution on how to better support victims of manipulated media. In the context of legality, there has been greater attention to deepfakes, including public consultations, law reform efforts, media attention, and other measures, but Henry et al. argue that there still exist several challenges and barriers to effectively aid victims of IBSA in the criminal justice space [11]. They suggest implementing broader measures, such as strengthening privacy laws, adjusting consumer platforms to make reporting more accessible, and raising awareness in prevention programs. Incorporating existing reform and policy recommendations gives further insight into understanding diverse victim experiences and their desire to not only recover but also seek justice.

2.3 Increasing Accessibility of Deepfakes

We specifically focus on the perspectives of laypeople because deepfakes are becoming increasingly prevalent for an increasingly large and diverse range of targets. This increased prevalence is largely due to the fact that deepfake generation technology is becoming more accessible and open to contribution. Tools such as FakeApp, FaceSwap, and ZAO allow people without any background in computer science to generate deepfakes quickly, and the open-source nature of many tools increases the speed of the technology's advancement [14]. Winter et al. focus on GitHub as a culprit behind the spread of deepfake generation technology [21]. Work by Westurlund et al. again acknowledges the increasing accessibility of deepfake generation, but more specifically describes the increased variety of types of individuals who are now able to make manipulated media because of this [20]. While our paper does not directly discuss the increasing accessibility of deepfake generation, we will focus on the consequential increased pool of laypeople victims and targets and their specific perspectives. A paper by Newton et al. explores GitHub as a source for deepfake generation software, identifies themes that suggest deepfakes serve as a source of toxic geek masculinity, and accordingly suggests approaches to harm mitigation [16]. We focus on the perceived harms of hypothetical victims or targets of manipulated media as opposed to the producers, so the suggestions for harm mitigation that we propose will be informed by the victims instead of the producers.

3 Methods

To answer our research questions, we conducted semi-structured interviews in November and December of 2021. Interviews lasted approximately 30 minutes and were conducted either over video call or in person. Due to time limitations, we interviewed only six people, all of whom were acquaintances of the researchers. We recruited these participants by reaching out to them directly. Prior to being interviewed, participants completed an online consent form, followed by a brief demographic survey. At the start of interviews, we asked participants whether they had any questions about the consent form. No compensation was provided.

This study was approved by Michelle Mazurek, acting with permission from the University of Maryland Institutional Review Board.

3.1 Interview Procedure

Our full interview protocol is in Appendix A. We took a semi-structured approach to our interview, which means that we generally followed the protocol but sometimes asked follow-up questions or skipped questions that had already been discussed.

We centered our interview around manipulated videos and photos, which we defined for participants as videos and photos that were created or changed in a significant way, using a computer, to depict someone in a situation that they were not actually in; we also specified that they should appear at least somewhat realistic. We began the interview by asking participants basic questions about their experience and attitudes regarding manipulated videos and photos, such as whether they had ever been depicted in one without their permission, and what concerns they had about negative consequences that could result from such a situation. We then asked questions about awareness and understanding of tools for manipulating videos and photos, such as deepfakes.

Next, we focused on deepfakes specifically. We defined deepfakes for participants as videos that have been manipulated using artificial intelligence, where the use of artificial intelligence makes it possible for a person without professional training to alter these videos realistically without editing each frame individually, just following instructions that are publicly available on the Internet. We then led participants through five exercises involving deepfakes: we showed two videos to demonstrate the capabilities of deepfakes and present a common narrative centered on making deepfakes of public figures for disinformation, and we narrated three hypothetical scenarios prompting participants to explore concerns related to interpersonal conflict, career and schooling, and sexually explicit material and bodily autonomy. After each, we asked them to talk about their reactions, such as feelings of surprise and new concerns.

Finally, we asked participants to imagine what information and resources they would want to protect themselves from negative consequences resulting from manipulated videos and photos. Due to the sensitive nature of this topic, we made sure to check in on participants' well-being before wrapping up the interview, and we described efforts to mitigate harms caused by deepfakes.

3.2 Analysis

Interviews were audio-recorded and transcribed using Otter.ai, an automated tool. We reviewed and corrected transcripts manually. All three researchers then coded three interviews together, developing a new codebook from scratch. We split up the remaining three interviews, which were each coded by one researcher alone. We communicated all updates to the codebook in order to facilitate consistent coding. Finally, we used the online tool Miro, which allowed us to visualize our codes as virtual sticky notes, to help identify patterns and generate themes.

3.3 Limitations

We asked participants to reflect on their concerns about manipulated videos and photos, but this may have led them to express or exaggerate concern, out of a desire to confirm our perceived hypotheses or to avoid answering negatively. This effect may be even stronger than usual because our participants knew their interviewer. We attempted to reduce this effect by encouraging participants at the start of the interview to speak freely and honestly. This was also aiming to reduce demand effects, which we further controlled by asking participants whether their concerns were new as of the interview.

We also acknowledge that participants' expressed concerns are significantly influenced by the deepfakes and hypothetical scenarios that we presented. While we intentionally chose examples that cover a variety of circumstances, they nonetheless represent only a small subset. Considering all of these threats to validity, we approach our analysis with the understanding that participants' expressed attitudes might not reflect the real world accurately. Instead, we focus more on our participants' awareness and understanding of facts; on their underlying assumptions and beliefs about things such as risk factors; and on the reasons they give for their expressed attitudes and changes in attitude.

4 Results

In this section, we describe our results, starting with our participants' demographics. After that, we discuss what participants know about manipulated media, what they believe, what their concerns are, how their attitudes changed, and what protective measures they desire.

4.1 Demographics

Table 1 summarizes our participant demographics. We observe that this is not a representative sample. In particular, ages ranged from 23 to 33, and all participants had received or were pursuing a bachelor's, graduate, or professional degree.

Participant	Age	Gender	Race	Education (degree)
P1	23	Female	White	Graduate or Professional
P2	23	Female	White	Graduate or Professional
P3	24	Female	Asian	Bachelor's
P4	33	Male	Asian	Bachelor's
P5	24	Female	Asian	Bachelor's
P6	30	Male	Asian	Graduate or Professional

Table 1: Participant demographics.

4.2 Knowledge and Experience

Familiarity with deepfakes. All six of our participants stated that they had previously heard of deepfakes; as with demographics, we note that this limits the scope of our findings. Some participants gave examples of deepfake subjects, including famous people and politicians; both P1 and P3 mentioned having seen deepfakes of former President Barack Obama specifically. P6 did not indicate knowledge of existing sexually explicit deepfakes, but they referred to revenge porn and expressed a concern that deepfakes could be used to doctor similar videos. Some participants also mentioned specific software that could be used to create deepfakes: P6 mentioned DeepFaceLab, a dedicated deepfake creation software, and P5 mentioned Snapchat and TikTok as social media apps with deepfake-like capabilities.

Four participants explicitly mentioned not understanding how deepfakes technically work or what software is used to create them. However, we found that many participants were able to describe applications of deepfakes: swapping faces, generating lip-sync videos from audio of speech, changing what a person is saying, changing the movement of a person's body, and stitching video clips together. In general, these do correctly describe common applications of deepfakes; stitching video clips together is not, to our knowledge, an existing type of deepfake, but it does bear a close resemblance to a technique that generates seamless lip-synced video by combining audio clips [19]. A few participants expressed a great deal of uncertainty in describing what deepfakes do, but by and large they still had accurate impressions. For example, P3 said, "Yeah, *I'm not too sure how it works.* I think, like my impression is like, you take like a video of someone speaking, for example, and *maybe* you can like move the mouth slightly to match a certain audio. But *I'm not too familiar* with the software" (italics added). Similarly, P5 said, "*I don't know* if it's manipulating like existing videos to put like someone else's face on it . . . *I don't know if I have the best understanding* of how it's actually done."

Lack of personal experience. No participant reported having been depicted without their permission in a manipulated video or photo. In addition, no participant reported having used deepfake software personally.

4.3 Assumptions and Beliefs

Assessing risk. After narrating the three hypothetical scenarios, we asked participants how likely they believed they were to be depicted in a similar deepfake; at other points throughout the interview, several participants also expressed beliefs about their likelihood of being depicted in a deepfake. Uniformly, participants said that they found it unlikely that they would be depicted in a deepfake without their permission. All participants did suggest personal characteristics that might increase risk, which did not apply to them. P6 mentioned being in a more prominent or managerial position as a risk factor; P1 mentioned having enemies or explaining their views on controversial topics; and P5 said that they "don't have a very big like presence on social media or like videos of me."

Requirements for creating a deepfake. Several participants expressed or implied beliefs about how difficult it is to create manipulated media. For example, P3 stated that "editing photos and videos takes time and skill." Regarding deepfakes in particular, P1 said that it is "really difficult" to make a deepfake, while P3 and P4 implied similar sentiments after we showed them the first deepfake, both saying that they might be surprised if they learned that it was easy to make. However, some participants expressed beliefs that manipulated media is accessible or becoming more so. P2 expressed concern about "when technology increases" and becomes more readily available, and P4 speculated that paying someone to create a short deepfake would probably not cost much.

Participants mentioned assumptions about what source material is required to create a deepfake. For example, P6 said, "for a realistic deepfake, you need thousands of photo references," and P3 said, "I'm assuming deepfakes require some video footage of the person that's in the deepfake." Interestingly, this is not entirely true; while common face swap and lip-sync techniques do rely on or improve with more source material [1, 19], there are also realistic deepfake techniques that require only a single source image [18].

4.4 Concerns and Perceived Harms

We found that participants had a diverse range of levels of concern regarding the consequences that deepfakes may have in their lives, ranging from little or no concern to high concern. In this section, we describe the major themes of specific concerns that participants expressed either in terms of themselves or someone they know personally being the target of a deepfake.

Professional, social, and self-esteem issues. All participants except for P6 described career or academic consequences caused by deepfakes as concerning. For example, P2 said, "I think it could get you in trouble with the department. It could affect your funding. It could affect your academic, like, record." All participants mentioned concerns relating to social consequences, which included drama or conflict in interpersonal or romantic relationships, with loss of trust being a particularly common concern. There was also a variety of concerns related to personal image or self-esteem, with P2 and P4 mentioning harassment, and all participants expect for P1 mentioning harm to personal image or reputation.

Mitigating harms. A few participants stated a belief that repercussions of being depicted in a deepfake could probably be reduced, but avoiding the repercussions would be a difficult task that could cause significant inconvenience. Most of these concerns centered on the difficulty of proving that a video is a deepfake. P1, multiple times, described the process of trying to explain the situation as potentially time-consuming. Participants also described concerns that they would not know how to report an incident to an app wherein it occurred, or how to remove a deepfake and its traces from the Internet. P5 anticipated that the whole process would be so stressful and harmful to their mental health and that proving that a video is a deepfake would be so difficult that they concluded, "I feel like the more likely thing in this scenario would just be for whatever repercussions to happen."

Lack of concern. P6 was the sole participant to claim that they were unconcerned about potential harms that could result from deepfakes. They trusted that deepfakes can be detected, either with or without the assistance of various tools. They also did not believe that their image is present enough online or concerning enough to be utilized in a deepfake, and they believe that any falsified content about them would be insignificant compared to the overwhelming total amount of content on the Internet.

4.5 Changes in Attitude

Throughout the interviews, we inquired about whether or not participants' concerns about deepfakes were changing. Half of the responses indicated a change in concern, always an increase, and half did not. The main reason cited for increasing concern was that our exercises increased participants' awareness of the capabilities and applications of deepfakes. The responses that expressed no change in concern were consistently explained by an existing familiarity with deepfakes or with the hypothetical scenario being described, or by a lack of surprise due to the expectation that technology in general is constantly advancing. All participants indicated that their level of concern was dependent on situational factors, including the importance or seriousness of the circumstances, the prominence of a deepfake subject's position, the amount of animosity that the creator of the deepfake has towards the subject, and the amount of time deepfake technology has had to advance compared with the amount of public awareness of deepfakes.

4.6 Wants, Needs, and Solutions

Towards the end of the interviews, we asked participants to describe any information or resources that they wanted to protect themselves from negative consequences of deepfakes. The general consensus among all participants was that it would be helpful to have more information about deepfakes. This included more information in general, such as with P2, who wasn't sure how to protect themself without more information. P2 and P4 both wanted better tools for protecting their privacy. All participants desired tools or guides to prevent, identify, or prove deepfakes, and P2 and P5 specified that these tools should be accessible.

5 Recommendations

For policymakers, technologists, and others who might provide support to victims of non-consensual manipulated media, we recommend that they consider providing support proactively. Our participants expressed that they considered themselves unlikely to be depicted in deepfakes, and some mentioned not knowing how to detect whether they had been portrayed in a manipulated video or photo. Given this, it is possible that non-consensual deepfakes may go undetected, even while they cause harms to the subject, such as damage to their reputation. Laypeople might benefit from assistance that does not require them to seek it out or even know that they have been depicted in a deepfake.

For journalists and others trying to raise awareness about deepfakes, we recommend that care be taken to describe deepfakes as an evolving group of diverse software and techniques, rather than as a single monolithic technology. Our study showed that that some people hold misconceptions, such as realistic deepfakes requiring video or lots of photos of the subject as source material, that are true for some common deepfakes but not for all. We recommend the same for techniques to detect deepfakes. One of our participants expressed confidence that forensics tools exist that can "ultimately" decide whether a video is a deepfake or not; while researchers are working on this, it is not clear that these techniques work consistently now or that they will in the future [13]. If people have more comprehensive and accurate mental models on their vulnerability to deepfakes, then they will be less likely to be lulled into a false sense of security, and more able to avoid unnecessary and burdensome precautionary measures.

6 Conclusion

We conducted semi-structured interviews to understand people's knowledge and attitudes about deepfakes and other forms of manipulated media. While this is a small and demographically skewed sample, we believe we have gleaned useful details about our participants' mental models and about the extent and limitations of their technical understanding. We intend to continue this work, alongside interviewing actual victims of non-consensual manipulated media, in the hopes of improving the security, privacy, and well-being of people who are affected or at risk.

Acknowledgments

We are thankful to Professor Michelle Mazurek and Kelsey Fulton for providing guidance throughout this project. We are also grateful to our participants, who gave us their time for no compensation.

References

[1] Sally Adee. What Are Deepfakes and How Are They Created? *IEEE Spectrum*, 2020.

[2] Henry Ajder, Giorgio Patrini, and Francesco Cavalli. Automating Image Abuse: Deepfake Bots on Telegram. Technical report, Sensity, October 2020.

[3] Samantha Bates. Revenge Porn and Mental Health: A Qualitative Analysis of the Mental Health Effects of Revenge Porn on Female Survivors. *Feminist Criminology*, 12(1):22–42, 2017.

[4] Ali Breland. Lawmakers Worry about Rise of Fake Video Technology. *The Hill*, 2018.

[5] Jacquelyn Burkell and Chandell Gosse. Nothing new here: Emphasizing the social and cultural context of deepfakes. *First Monday*, 24(12), Dec. 2019.

[6] Samantha Cole. AI-Assisted Fake Porn Is Here and We're All Fucked. *Motherboard*, December 2017.

[7] Jesselyn Cook. Here's what it's like to see yourself in a deepfake porn video. *The Huffington Post*, 2019.

[8] Adrienne de Ruiter. The distinct wrong of deepfakes. *Philosophy & Technology*, 34, Dec. 2021.

[9] Mollie C. DiTullio and Mackenzie M. Sullivan. A Feminist-Informed Narrative Approach: Treating Clients Who Have Experienced Image-Based Sexual Abuse. *Journal of Feminist Family Therapy*, 31(2-3):100–113, April 2019.

[10] Diana Freed, Sam Havron, Emily Tseng, Andrea Gallardo, Rahul Chatterjee, Thomas Ristenpart, and Nicola Dell. "Is my phone hacked?" Analyzing Clinical Computer Security Interventions with Survivors of Intimate Partner Violence. In *Proceedings of the ACM on Human-Computer Interaction*, volume 3, pages 202:1–202:24. ACM, 2019.

[11] Nicola Henry, Asher Flynn, and Anastasia Powell. Policing image-based sexual abuse: stakeholder perspectives. *Police Practice and Research*, 19:565–581, Nov. 2018.

[12] Nicola Henry, Asher Flynn, and Anastasia Powell. Image-based sexual abuse: Victims and perpetrators. *Trends & Issues in Crime and Criminal Justice*, (572):1–19, March 2019.

[13] Shehzeen Hussain, Paarth Neekhara, Malhar Jere, Farinaz Koushanfar, and Julian McAuley. Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples. pages 3348–3357, 2021.

[14] Momina Masood, Marriam Nawaz, Khalid Mahmood Malik, Ali Javed, and Aun Irtaza. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *CoRR*, abs/2103.00484, 2021.

[15] Clare McGlynn and Erika Rackley. Image-Based Sexual Abuse. *Oxford Journal of Legal Studies*, 37(3):534–561, September 2017.

[16] Olivia B. Newton and Mel Stanfill. My nsfw video has partial occlusion: deepfakes and the technological production of non-consensual pornography. *Porn Studies*, 7(4):398–414, 2020.

[17] Yanet Ruvalcaba and Asia A. Eaton. Nonconsensual pornography among U.S. adults: A sexual scripts framework on victimization, perpetration, and health correlates for women and men. *Psychology of Violence*, 10(1):68–78, 2020.

[18] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First Order Motion Model for Image Animation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[19] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing Obama: Learning Lip Sync from Audio. *ACM Transactions on Graphics*, 36(4):95:1–95:13, July 2017.

[20] Mika Westerlund. The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9:39–52, 11 2019.

[21] Rachel Winter and Anastasia Salter. Deepfakes: uncovering hardcore open source on github. *Porn Studies*, 7(4):382–397, 2020.

A Interview Protocol

A.1 Background

Q: Has anyone ever depicted you without your permission in a manipulated video or photo? When I say “manipulated,” I mean it was created or changed in a significant way, using a computer, to depict you in a situation that you were not actually in. The video or photo should also appear at least somewhat realistic.

[IF YES]

Q: Could you please describe, at a level that you’re comfortable with, what was depicted in this video or photo? Remember that you don’t have to answer a question if you’re uncomfortable doing so.

Q: How did you feel after learning about this video OR photo?

[IF NO]

Q: Have you heard of anyone being depicted without their permission in a manipulated video or photo? This doesn’t have to be someone you know.

Q: Are you concerned about potential negative consequences that might result from someone depicting you or someone you know in a manipulated video or photo without permission?

[IF YES]

Q: Could you describe some of those concerns to me?

Q: To your best ability, can you recall whether you had these concerns before this interview, or whether these are new concerns that you’re thinking of now?

[IF NO]

Q: Could you talk more about why you’re not concerned?

Q: Let’s suppose someone depicts you in a manipulated video or photo without your permission, in such a way that there might be negative consequences for you. In your opinion, who is most likely to do this, and why?

A.2 Knowledge

Q: Have you heard of software, such as Adobe Photoshop, that can be used by non-professionals to manipulate photos realistically?

[IF YES]

Q: In your understanding, how can this kind of software help someone manipulate photos without professional training?

Q: Have you ever used any of this software before?

[IF NO]

Q: In your understanding, how can computers help someone manipulate photos without professional training?

Q: Have you heard of techniques, such as deepfakes, that can be used by non-professionals to manipulate videos realistically?

[IF YES]

Q: In your understanding, how can this kind of technique help someone manipulate videos without professional training?

Q: Have you ever used any of this software before?

[IF NO]

Q: In your understanding, how can computers help someone manipulate videos without professional training?

A.3 Clips and Scenarios

(1) Now I’d like to show a deepfake video clip. Deepfakes are videos that have been manipulated using artificial intelligence. The use of artificial intelligence makes it possible for a person without professional training to alter these videos realistically without editing each frame individually, just following instructions that are publicly available on the Internet. The video I am about to show is a character’s speech taken from the movie Iron Man 2; the original footage with the original actor, Robert Downey Jr., will be on the left. The deepfake, on the right, is manipulated so that the character appears to have the face of Tom Cruise, a different actor. The original sound has also been replaced by a voice actor to sound like Tom Cruise.

Q: Were you surprised in any way by this deepfake video? Why (why not)?

Q: Do you have any new or changed concerns about someone depicting you or someone you know in a manipulated video or photo without permission?

(2) Now, I’m going to show a deepfake video clip that used an audio recording of actor Jordan Peele speaking in order to generate a fake video of former President Barack Obama saying the same words. This video includes one instance of strong language; are you okay with that?

Q: Were you surprised in any way by this deepfake video? Why (why not)?

Q: Do you have any new or changed concerns about someone depicting you or someone you know in a manipulated video or photo without permission?

(3) Now, let's consider a hypothetical scenario. You receive a message from someone you know. The message includes a video clip depicting one of your friends saying something offensive about another friend. This video is actually a deepfake, but it is presented as if it were real.

Q: What kinds of negative consequences do you think might result from this?

Q: How likely do you think it is that someone might create a deepfake video like this one depicting you?

Q: How concerned would you be about negative consequences if someone were to create a deepfake video like this one depicting you?

(4) Let's consider another hypothetical scenario. An administrator at work or at school contacts you. One of your peers anonymously emailed them a video clip depicting you violating your workplace or school's policy; for example, stealing property or plagiarizing someone else's work. This video is actually a deepfake, but it is presented as if it were real.

Q: What kinds of negative consequences do you think might result from this?

Q: How likely do you think it is that someone might create a deepfake video like this one depicting you?

Q: How concerned would you be about negative consequences if someone were to create a deepfake video like this one depicting you?

(5) Finally, let's consider one more hypothetical scenario. Someone you know shows you a dating app profile featuring photos and a video that appear to depict one of your friends, who is in a long-term relationship. In the video, your friend appears to be partially nude. This profile was actually created by someone else; the photos are actual photos of your friend, but the video is a deepfake created from a stranger's partially nude video.

Q: What kinds of negative consequences do you think might result from this?

Q: How likely do you think it is that someone might create a deepfake video like this one depicting you?

Q: How concerned would you be about negative consequences if someone were to create a deepfake video like this one depicting you?

Q: After considering these hypothetical scenarios, do you have any new or changed concerns about someone depicting you or someone you know in a manipulated video or photo without permission?

A.4 Exploring Solutions

Q: Is there any information that you wish you knew in order to protect yourself from negative consequences that could result from manipulated videos and photos?

Q: Are there any resources that you wish you had in order to protect yourself from harms that could result from manipulated videos and photos? Feel free to be creative. Resources could be people, how-to guides, tools, or anything you can think of, and don't worry about whether or not they currently exist or are available.

Q: Do you have any final thoughts? Is there anything else you think we should know?

A.5 Debrief

Q: Would you like to hear more about efforts to mitigate the negative consequences of deepfakes?

[IF YES]

(1) There are lots of people, including technologists, lawyers, and policymakers, who are working on better solutions to detect and regulate deepfakes, increase awareness of deepfakes, and generally protect people from negative consequences.

(2) We've dealt with similar problems before; since photographs were invented, there have been deceptive altered photos. Society didn't collapse—people adjusted their expectations for how much photos could be trusted and continued on. Deepfake videos are concerning, for sure, but we will find ways of adjusting.

Q: Do you have any questions for me about this research?

Q: Is there anything you wish we had done differently, regarding this interview?