CYBERPSYCHOLOGY, BEHAVIOR, AND SOCIAL NETWORKING
Volume 24, Number 3, 2021
© Mary Ann Liebert, Inc.
DOI: 10.1089/cyber.2021.29208.jth

GUEST EDITORIAL

The Social Impact of Deepfakes

Jeffrey T. Hancock, PhD and Jeremy N. Bailenson, PhD

IN MANY WAYS, this special issue was inspired by a visit to the University of Washington in 2018. Seitz and his colleagues had just published the algorithms[1] that enabled their now famous Obama video, in which a few hours of simple audio clips could drive a high-quality video lip syncing. At the end of the video, a young Obama audio clip is parroted perfectly by a video version of Obama who is twice his age. This is likely the most canonical, if not the original ''deepfake'' video. It is enabled by machine learning, which uses multiple videos as a training set to categorize speech into ''mouth shapes,'' which are then integrated into an existing target video. The outcome is a stunningly real video that few would give a second glance to—it simply looks like President Obama talking. Aside from the realism of the videos, there were two striking things about Seitz's presentation.

First, the algorithms that build deepfakes are easier to build than detect, based on the very nature of the Generative Adversarial Networks employed according to Goodfellow,[2] these models are constructed by pitting ''counterfeiters'' against ''police,'' and successful models by definition have already shown that the fake can beat detection methods. Indeed, since deepfakes have migrated from top computer science laboratories to cheap software platforms all over the world, researchers are also focusing on defensive algorithms that could detect the deception (see Tolosana et al.,[3] for a recent review). But Seitz was not confident about this strategy, and likened the spiral of deception and detection with an arms race, with the algorithms that deceive having the early advantage compared with those that detect.

The second eye opener was the many social and psychological questions that these deepfakes raised: does exposure to deepfakes undermine trust in the media? How might deepfakes be used during social interactions? Are there strategies for debunking or countering deepfakes? There has been ample work done in computer science on automatic generation and detection of deepfakes, but to date there have only been a handful of social scientists who have examined the social impact of the technology. It is time to understand the possible effects deepfakes might have on people, and how psychological and media theories apply.

A Scarcity of Empirical Research

At the time of this writing, only a few studies have examined the social impact of deepfakes,[4] despite the popularity of face swap platforms (e.g., the Zao app[5]). This special issue sought out the first generation of deepfake research that examines the psychological, social, and policy implications of a world in which people can easily produce and disseminate videos of events that never actually occurred, but that are indistinguishable from real videos.

Although there have been dozens of studies looking at false memory acquisition and social influence from altered still images (i.e., Garry and Wade[6]), the psychological processes and consequences of viewing artificial intelligence (AI)-modified video remain largely unstudied. Surprisingly, the best starting point for understanding the impact of deepfakes is immersive virtual reality (VR). In VR, one can build ''doppelgangers,'' three-dimensional (3D) models of a given person, based on photogrammetry and other techniques that create a 3D structure from a series of two-dimensional (2D) images. Once the doppelganger is built, it is simple to apply stock animations onto the 3D models and then show the people the VR deepfake scene, either in a head-mounted display or rendered as a normal 2D video animation. VR deepfakes are impactful. Compared with watching scenes of another person, watching your own doppelganger causes encoding of false memories in which participants believe they actually performed the deepfake activity,[7] more exercise behavior after watching a positive health outcome,[8] and brand preference for products used by the virtual self in the deepfake.[9] These studies all involve relatively cartoonish animations, and although some of the mechanisms at play in these previous findings may still be relevant (e.g., building self efficacy from watching the self succeed), it is highly probable that new psychological mechanisms and outcomes are at play when a deepfake video is perceptually indistinguishable from a real video.

Some Insights from Deception Research

At the core of deepfakes is deception, which involves intentionally, knowingly, and/or purposely misleading another person.[10] The deception detection literature suggests that people are not particularly good at detecting deception when assessing messages and can relatively easily acquire false beliefs. Meta-analyses of deception detection studies suggest that people perform only slightly above chance when evaluating a message as either true or deceptive.[11] Importantly, this level of accuracy is not affected by the medium in which the message is conveyed. Studies have shown that deception detection is approximately the same whether the message is conveyed through text (e.g., a court transcript, an Internet chat log), an audio recording (e.g., a voicemail, a radio program), or a video (e.g., an interrogation video).[12] Although this may seem surprising given the richer detail available in video, accuracy tends to be at chance regardless of medium because there are no reliable signals to human deception (i.e., there is no Pinocchio's nose) and we tend to trust what others say.[10] But the vast majority of research on video-based deception has examined the verbal content of speech, for example, a person telling a lie, as opposed to the movement and/or form of a person's body. One of the most exciting aspects of this special issue is to explore deception that is not based solely on lies told with words, but instead the complete fabrication of verbal *and* nonverbal behaviors.

Although the rates of detection are likely similar to other media, the *impact* of deception by deepfake has the potential to be greater than that of verbal deception because of the primacy of visual communication for human cognition. Deepfakes not only change verbal content, but they also change the visual properties of how the message was conveyed, whether this includes the movement of a person's mouth saying something that he or she actually did not, or the behavior of a person doing something that he or she did not. The dominance of visual signals in human perception is well established.[13] For example, under many circumstances, humans rely more on visual information than other forms of sensory information, a phenomenon referred to as the Colavita visual dominance effect.[14] In the basic Colavita paradigm, participants have to make speeded responses to a random series of auditory, visual, or audiovisual stimuli. Participants are instructed to make one response to an auditory target, another response to a visual target, and to make both responses whenever the auditory and visual targets are presented at the same time. Participants have no problem in responding to the audio and video targets separately, but when they are presented together, they often fail to respond to the auditory targets. It is as if the visual stimuli extinguish the audio stimuli. Following the communication literature, people are also more likely to recall visual messages than verbal messages,[15,16] and misleading visual information is more likely to generate false perceptions than misleading verbal content because of the ''realism heuristic,'' in which people are more likely to trust audiovisual modalities over verbal because the content has a higher resemblance to the real world.[17]

Video was the last frontier—the one medium consumers could watch and not automatically assume it could be faked. But what happens when we learn that video can be ''photoshopped'' as easily as images can be? Can we believe any media that we see? The philosopher Don Fallis[18] refers to this as the *epistemic threat* of deepfakes. His argument flows from the power of visual media to carry information, which refers to how much signal is conveyed by a message. Because of the dominance of the visual system, videos have high information carrying potential—that is, we tend to believe what we see in a video, and as a result videos have become the ''gold standard'' of truth. But as deepfakes proliferate and awareness that videos can be faked spread through the population, the amount of information that videos carry to viewers is diminished. Even if a video is genuine and a viewer would acquire true beliefs, distrust born of deepfakes would prevent a person from actually believing what they saw. The epistemic threat for Fallis is that deepfakes will interfere with our ability to acquire knowledge about the world by watching media. The implications for our shared understanding of the world, and the role that journalism and other media play in constructing that world, may be seriously undermined.

Deepfake Consequences

Unfortunately, one of the few empirical studies on deepfakes provides some early evidence that worryingly bears this philosophical account out. In a study looking at the effect of deepfakes on trust in the news, Vaccari and Chadwick[19] found that although people were unlikely to be completely misled by a deepfake (at least with the technology they were using), exposure to the deepfake increased their uncertainty about media in general. Confirming the worst expectations, that sense of uncertainty led participants to reduce their trust in news, much as Fallis's account of epistemic threat predicts.

Deepfakes also have interpersonal consequences. As the VR studies already described suggest, video deepfakes have the potential to modify our memories and even implant false memories, and they can also modify a person's attitudes toward the target of the deepfake. One recent study revealed that exposure to a deepfake depicting a political figure significantly worsened participants' attitudes toward that politician.[20] Even more worryingly, given social media's ability to target content to specific political or demographic groups, the study revealed that microtargeting the deepfake to groups most likely to be offended (e.g., Christians) amplified this effect relative to sharing the deepfake with a general population.

Although these implications paint a discouraging portrait of a future with deepfake technology, this take assumes a relatively passive consumer of media. It is important to recall that humans have been adapting to novel forms of deception for millenia.[21] People tend to be trusting of one another until they have some reason to become suspicious or more vigilant, a state that Levine[10] refers to as a trust default. We move out of our trust default when we learn about inconsistent information, or we are warned by a third party, or we are educated about novel deceptive techniques. For example, email spam is much less effective than when it first emerged, in part because people are aware of it.

In the same way, it is possible for people to develop resilience to novel forms of deception such as deepfakes. For example, advertising frequently relies on misleading visual information (e.g., drink this beer, have beautiful friends; smoke this cigarette, experience the great outdoors). Over time, consumers get their guard up and are not fooled by advertising, in part because they develop a schema of expectations for advertising.[22] Indeed, we develop expectations like this for most media we consume. For example, deepfake technology is already used in Hollywood movies, for example, the portrayal of Princess Leia in Star Wars VIII after the actor Carrie Fisher had died. Most people discount the deepfake as fiction given that they are watching a fictional movie. But, an important question is whether the visual evidence starts to chip away at the viewers' memory that the actress has passed away, regardless of their knowledge that it is a movie? This special issue helps to assimilate the theories and methods that begin to answer these questions.

An important harm we have not yet considered is the nonconsensual victim portrayed in a deepfake to be doing or saying something that they did not. One of the most common early forms of deepfakes is the alteration of pornography, depicting nonconsensual individuals engaging in a sex act that never occurred typically by placing a person's face on another person's body. Given the power of the visual system in altering our beliefs already described, and the influence that such deepfakes can have on self identity, the impact on a victim's life can be devastating. Although empirical research to date is limited, it is not difficult to imagine how deepfakes could be used to extort, humiliate, or harass victims.

The First Wave of Empirical Work on Deepfakes

The articles in the rest of this special issue represent different levels of analysis and a diverse range of empirical methods, including qualitative analyses, surveys, experiments, and policy analysis. They also represent international diversity, with articles from Europe, Asia, and North America. In our view, the present articles are part of the very first wave of empirical work on the social impacts of deepfakes.

One set of articles touches on the current state of deepfakes. The first article provides a historical snapshot of the 10 most popular current deepfakes on Youtube and analyzes linguistic responses through comments from viewers (see Lee et al.). The second article mines Reddit in 2018 to gauge the climate surrounding deepfakes, and uses those data to innovate possible solutions to adverse use cases (see Brooks). The third article (see Cochran and Napshin) surveys students to gauge their awareness and concerns about deepfakes, and the degree to which platforms are responsible for regulating the technology.

The next group of articles provides some initial insights into some of the psychological dynamics of deepfakes on self perception. In the first study, authors (see Wu et al.) examine how young women evaluate their own appearance before and after an exposure to a deepfaked image that blended an image of themselves with a celebrity. Contrary to intuitive predictions, they demonstrated positive effects of viewing oneself within a deepfake, and provide mechanisms for how deepfakes influence self perception. This study paves the way for others to study the possible prosocial applications of deepfakes. In the next empirical study (see Weisman and Peña), the authors investigate how exposure to a reconstructed version of the self or ''talking head'' created by an AI program influences trust toward AIs. They find that exposure to a talking head with the participant's face reduced affect-based trust toward AIs, and that because they used software that produced strange artifacts on the eyes of the faces, uncanny valley perceptions mediated this effect.

Two of the articles evaluate the effectiveness of strategies for combating the impact of deepfakes and provide some important avenues for protecting populations against the new technology. In what is likely the first study evaluating a media literacy program targeting deepfakes (see Hwang et al.), the authors show efficacy for a media literacy program in Korea that builds resilience to believing deepfakes. The second article (see Iacobucci et al.) examines the roles that priming and individual differences play in participants' ability to detect deepfakes. For example, they find that people with a proclivity to believe false information are indeed more susceptible to believing deepfakes, but they also provide some innovative priming ideas for resisting deepfakes.

In the final article in the special issue, authors Vasileia and Aalia consider how all these previously discussed issues translate into policy. The authors build out a framework by conducting a case study review of Canadian Policy to identify the threats associated with deepfake pornography and provide much needed lessons for policy development. This study builds on other research currently laying out the implications for policy of deepfakes.[23]

Future Research

In this special issue, we urge researchers to begin to study the social issues surrounding deepfake technology. The studies in this volume do a fantastic job of mapping out the research questions, applying theory to the phenomenon, and creating new tools to apply to future research. But this study is preliminary, and we urge scholars to build upon this study as deepfake use continues to grow.

However, there is another, and related, frontier that needs attention. Currently, when we discuss deepfakes, we are referring to recorded video. But machine learning has advanced sufficiently to enable real-time deepfakes: AI-powered filters. These filters allow for modifying or optimizing the video content of a videoconference in real time, such as making a person's eye gaze appear as though it is aimed at the camera even though it is pointed elsewhere on the screen. In addition to managing joint attention in awkward video settings, other deepfake filters are being developed to optimize for other interpersonal dynamics, such as warmth or interpersonal attraction. For example, Oh et al.[24] employed a real-time filter to enhance the amount of smiling in dyads, and demonstrated that partners in the enhanced smiling conditions felt more positively after the conversation, and actually used more positive words during their conversation based on linguistic analysis. It is critical to note that these downstream effects occurred even though the participants were not aware, and almost never detected, of the smiling filter. Researchers at the MIT Media laboratory are developing ''personalized role models'' using deepfake technology to alter a real-time video stream to allow speakers to see versions of themselves excelling at speaking tasks in a confident manner, and are demonstrating effects not only on mood but also on task creativity.[25] This use of AI to modify one's self presentation during videoconferencing is a form of AI-mediated communication, which refers to ''interpersonal communication in which an intelligent agent operates on behalf of a communicator by modifying, augmenting, or generating messages to accomplish communication goals.''[26]

Although deepfake technology has the potential to undermine our trust in media or falsely influence our beliefs about the world, it may also become more commonplace and mundane as people use deepfake technology to improve their day-to-day communication. As the mentioned discussion makes clear, and the articles in this special issue highlight, there are many important psychological, social, and ethical issues that require innovative and careful empirical analyses of the social impact of deepfake technologies.

References

1. Suwajanakorn S, Seitz SM, Kemelmacher-Shlizerman I. Synthesizing Obama: learning lip sync from audio. ACM Transactions on Graphics (ToG) 2017; 36:1–13.
2. Goodfellow I. Nips 2016 tutorial: Generative adversarial networks 2016; arXiv preprint arXiv:1701.00160.

3. Tolosana R, Vera-Rodriguez R, Fierrez J, et al. Deepfakes and beyond: a survey of face manipulation and fake detection. Information Fusion 2020; 64:131–148.
4. Ahmed S. Who inadvertently shares deepfakes? Analyzing the role of political interest, cognitive ability, and social network size. Telematics and Informatics 2021; 57:101508.
5. Doffman Z. Chinese deepfake app ZAO goes viral, privacy of millions 'at risk'. Forbes Magazine. 2019. https://www.forbes.com/sites/zakdoffman/2019/09/02/chinese-best-ever-deepfake-app-zao-sparks-huge-faceapp-like-privacy-storm/?sh=2951ebb88470 (accessed Jan. 27, 2021).
6. Garry M, Wade KA. Actually, a picture is worth less than 45 words: narratives produce more false memories than photographs do. Psychonomic Bulletin & Review 2005; 12: 359–366.
7. Segovia KY, Bailenson JN. Virtually true: children's acquisition of false memories in virtual reality. Media Psychology 2009; 12:371–393.
8. Fox J, Bailenson JN. Virtual self-modeling: the effects of vicarious reinforcement and identification on exercise behaviors. Media Psychology 2009; 12:1–25.
9. Ahn SJ, Bailenson J. Self-endorsed advertisements: when the self persuades the self. Journal of Marketing Theory and Practice 2014; 22:135–136.
10. Levine TR. (2019) *Duped: truth-default theory and the social science of lying and deception.* Tuscaloosa, AL: University Alabama Press.
11. Bond Jr CF, DePaulo BM. Accuracy of deception judgments. Personality and Social Psychology Review 2006; 10:214–234.
12. Hancock JT, Woodworth MT, Goorha S. See no evil: the effect of communication medium and motivation on deception detection. Group Decision Negotiation 2010; 19: 327–343.
13. Posner M, Nissen M, Klein R. Visual dominance: an information-processing account of its origins and significance. Psychological Review 1976; 83:157–171.
14. Koppen C, Spence C. Seeing the light: exploring the Colavita visual dominance effect. Experimental Brain Research 2007; 180:737–754.
15. Graber DA. Seeing is remembering: how visuals contribute to learning from television news. Journal of Communication 1990; 40:134–155.
16. Prior M. Visual political knowledge: a different road to competence? Journal of Politics 2013; 76:41–57.
17. Sundar S. (2008). The MAIN model: a heuristic approach to understanding technology effects on credibility. In Metzger M, Flanagin A, eds. *Digital media, youth, and credibility*. Cambridge, MA: MIT Press, pp. 73–100.
18. Fallis D. The epistemic threat of deepfakes. Philosophy & Technology 2020; 1–21.
19. Vaccari C, Chadwick A. Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social Media and Society 2020; 6:1–13.
20. Dobber T, Metoui N, Trilling D, et al. Do (microtargeted) deepfakes have real effects on political attitudes? The International Journal of Press/Politics. 2019; 26:69–91.
21. Reeves B, Nass C. (1996). *The media equation: how people treat computers, television, and new media like real people.* Cambridge, United Kingdom: Cambridge University Press.
22. Kim SJ, Hancock JT. How advertorials deactivate advertising schema: MTurk-based experiments to examine persuasion tactics and outcomes in health advertisements. Communication Research 2017; 44:1019–1045.
23. Liv N, Greenbaum D. Deepfakes and memory malleability: false memories in the service of fake news. AJOB Neuroscience 2020; 11:96–104.
24. Oh SY, Bailenson J, Krämer N, et al. Let the avatar brighten your smile: effects of enhancing facial expressions in virtual environments. PLoS One 2016; 11:e0161794.
25. Leong JS. Investigating the use of synthetic media and real-time virtual camera filters for supporting communication and creativity. Unpublished master's thesis, Massachusetts Institute of Technology, 2021.
26. Hancock JT, Naaman M, Levy K. AI-mediated communication: definition, research agenda, and ethical considerations. Journal of Computer-Mediated Communication 2020; 25:89–100.

Jeffrey T. Hancock, PhD and Jeremy N. Bailenson, PhD
Department of Communication
Stanford University
USA