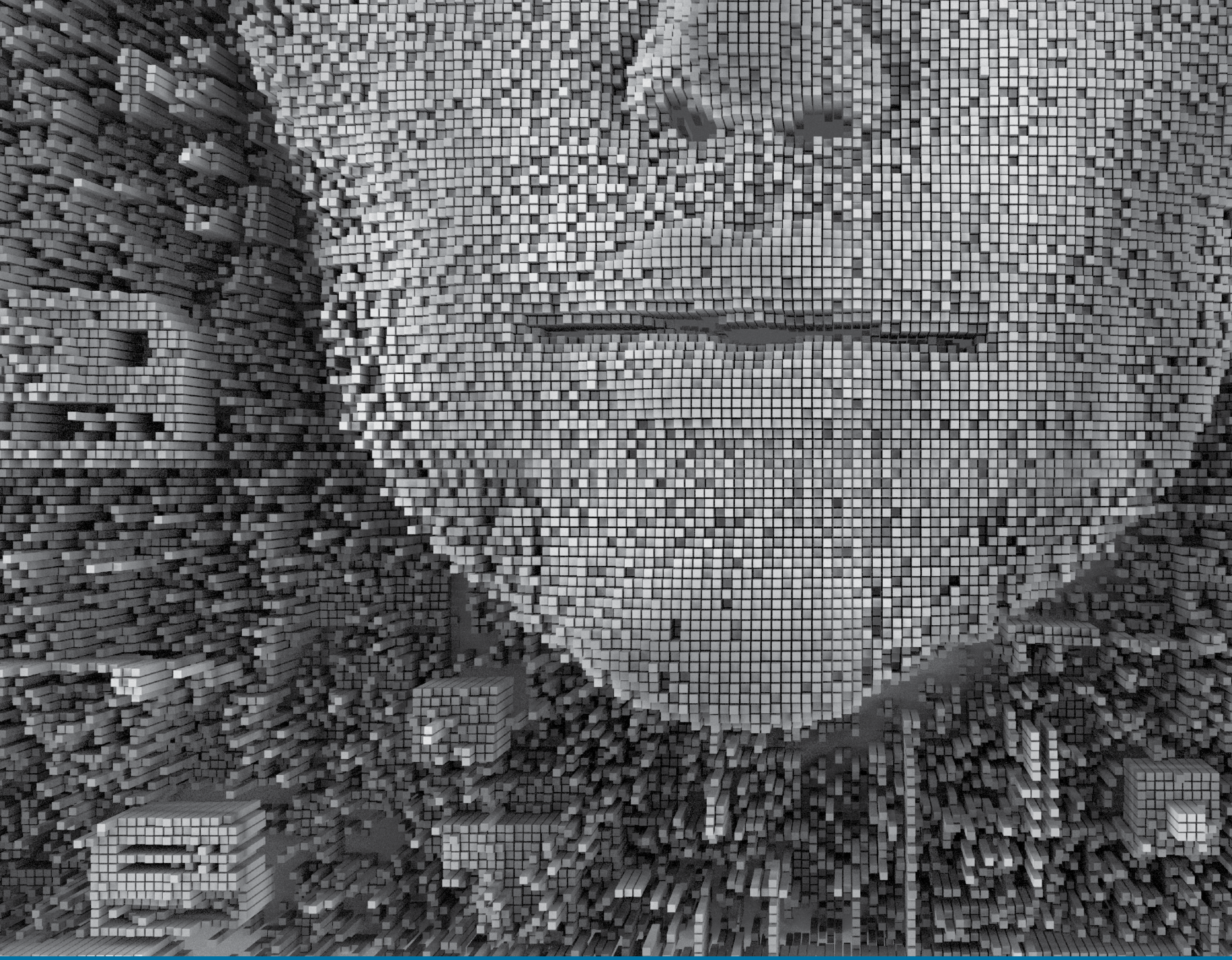

Disinformation and the "Deepfake"

Harith Khawaja and Christopher Koopman

This past May, a video of Mark Zuckerberg declaring that he owned billions of people's stolen data was posted on Instagram, a platform owned by Facebook.[1] The video was fake. Using complicated data processing methods to alter publicly available footage of Zuckerberg, researchers manipulated the video to put words in Zuckerberg's mouth. This is just the latest example of the new challenge facing social media platforms, users, and policymakers trying to understand how to sort fact from fiction on the Internet.

One concern is that these videos could portray individuals doing or saying things they never did in an effort to spark unjustified controversy. For example, presidential candidates or the President might appear in places they never were, engaging in illegal activities. Police officers may be depicted as shooting unarmed individuals while shouting slurs. Videos could show Muslims at a local mosque celebrating ISIS, an event that could possibly stoke violence against that community.[2] Or even doomsday situations in which newscasters announce the start of a nonexistent nuclear war.[3]

Another concern is that while these doctored videos are still easy to spot as fake now, it is becoming increasingly harder to do so. Standard video editing techniques can, at minimum, insert new objects, alter the pitch of people's voices, or change colors. The technology behind these "deepfakes" (the term given to these "fake" videos generated by algorithms called "deep" neural networks) allows programmers to superimpose faces and voices in real-time, or even generate entire videos or audio files from scratch.[4] As a result, the doctored content is often indistinguishable from an authentic video.

This has a two-fold implication. First, viewers are fooled into believing that false content is real. Second, with the proliferation of deepfakes, viewers may be less willing to believe in real content because they would simply dismiss it as a deepfake. The resulting atmosphere is one where people can no longer believe what they see.

While many of the concerns about deepfakes involve worries about the future, these fake videos are already affecting real people. Since 2017, fabricated pornographic videos with the faces of celebrities like Scarlett Johansson, Maisie Williams, Taylor Swift, Aubrey Plaza, and Gal Gadot have been created and uploaded to online platforms like Reddit.[5] Standalone apps have been released that enable users with no technical experience to create pornographic videos of people they know just by uploading a few photos.[6] In one recent case, a $50 application available for Windows and Linux machines called "DeepNude" allowed users to undress a photo of a woman with a single click.[7] After some backlash, the app was taken down. And critics point out that deepfakes have been repeatedly used to threaten, blackmail and slander women, and to establish dominance over their bodies, especially by representing them in non-consensual videos.[8]

These concerns may seem strong enough for policymakers to do something, but why hasn't anything been done about deepfakes? For one thing, to ban deepfakes is to ban the technology that's used to create them. The algorithmic basis for deepfakes can be assembled using open-source software toolkits developed and maintained by Google and Facebook, like Tensorflow and PyTorch. When it comes down to feeding these algorithms the data they need, programmers can obtain audio, video, and pictures online, for little to no cost. For reference, the first deepfake porn creators used Google image search, stock photos, and YouTube videos to train their algorithms.[9] As deepfake technology becomes more and more accessible, it becomes increasingly harder — perhaps impossible — to ban deepfakes altogether.

Banning deepfakes would also forgo the positive uses of the technology. The algorithms behind doctored videos have also been used to create language processing systems like Alexa and Siri, music in the spirit of Bach, and art that has been auctioned at

Christie's.[10] They have brought movie stars like Peter Cushing back from the dead to feature in film sequels.[11] And they are being used to generate high-resolution images to improve the accuracy of algorithms used in the healthcare industry.[12] Generated video could potentially be used in schools to teach history — imagine being transported back to World War I — and create images of extinct species that could promote conservation purposes. By banning deepfakes out of fear, we risk losing the benefits.

So how else can we effectively moderate how deepfakes are used? One suggestion has been to strip the legal immunity online platforms have under federal law.[13] By making platforms liable for user-posted content, the argument goes, platforms would be incentivized to remove harmful content like deepfake porn, which would make the online world safer for everyone.[14]

Yet, this argument may end up having more costs than benefits. The Electronic Frontier Foundation, for example, has described section 230 of the Communications Decency Act (which created the immunity that online platforms enjoy) as "perhaps the most influential law to protect the kind of innovation that has allowed the Internet to thrive since 1996."[15] The promise of immunity from liability has allowed Facebook, Twitter, YouTube, Yelp and other startups to take off, and is why the Internet ecosystem has been so dynamic and competitive. Amending this protection in the name of stifling deepfakes could deal a far-reaching blow to the Internet.

This is not to say that nothing could be done. And perhaps efforts should be less focused on banning and more on identifying deepfakes. Once a video is identified as such, efforts can then be made to inform viewers. This, however, is a difficult task.

Some argue that deepfakes can be spotted with the naked eye. By examining a video closely enough, and by concentrating on features like the perimeter of people's faces and background colors, experts can identify whether or not it is fake.[16] The idea is that doctored videos often have irregularities in color, sound, pixelation and content. By detecting these irregularities, we could expose deepfakes.

While this might be a successful short-term approach, it is not going to always work. As we mention above, the quality of deepfakes produced continues to improve. Over time, these irregularities will become less and less frequent. Experts have predicted that, within a year, deepfakes will become visually undetectable by humans.[17] Beyond that point, fake and real will become indistinguishable. This troublesome thought has prompted researchers to develop technologies that could do the identification for us. The US Department of Defense's Advanced Projects Research Agency (DARPA, which built the precursor to the modern internet), has spent millions of dollars toward this end. Their "media forensics" approach has endowed researchers to develop algorithms that can identify telltale signs of media manipulation much more accurately than the human eye.[18] Research into this approach is ongoing but shows promise; one experiment achieved up to 92% accuracy.[19]

Deepfake videos raise hard questions with no straightforward answers, especially related to how to regulate them. The first step for effective moderation, however, is increased public awareness. Policymakers should take the necessary steps to get themselves acquainted with the issues surrounding the virality of deepfakes and the immense personal and institutional threats they pose. On this front, there has been some activity. This summer, the US House of Representatives held the first hearing on deepfakes. While these efforts continue, it is important that we take a balanced approach that allows for the benefits of the technology used to create deepfakes to emerge while seeking to mitigate the harms that could occur as a result of fabricated content. Only then can we begin to provide real solutions to deepfakes.

Harith Khawaja is a Technology and Policy Intern with the Center for Growth and Opportunity. Christopher Koopman is the Senior Director of Strategy and Research for the Center for Growth and Opportunity and Adjunct Director of JMI's Center for Technology and Innovation.

References

1 Cole, "This Deepfake of Mark Zuckerberg Tests Facebook's Fake Video Policies."

2 Chesney and Citron, "Deep Fakes."

3 Cook, "Deepfake Videos And The Threat Of Not Knowing What's Real."

4 Chesney and Citron, "Deep Fakes."

5 Cole, "AI-Assisted Fake Porn Is Here and We're All Fucked - VICE."

6 Cole, "We Are Truly Fucked."

7 Cole, "This Horrifying App Undresses a Photo of Any Woman With a Single Click."

8 Cole, "Deepfakes Were Created As a Way to Own Women's Bodies—We Can't Forget That."

9 Cole, "AI-Assisted Fake Porn Is Here and We're All Fucked - VICE."

10 "Is Artificial Intelligence Set to Become Art's next Medium?"

11 Itzkoff, "How 'Rogue One' Brought Back Familiar Faces."

12 Hao, "A New Way to Use the AI behind Deepfakes Could Improve Cancer Diagnosis."

13 47 USC § 230 (2019).

14 Cook, "Here's What It's Like To See Yourself In A Deepfake Porn Video."

15 "Section 230 of the Communications Decency Act."

16 Hu, "It Is Your Duty to Learn How to Spot Deepfake Videos."

17 Summerville, "'Deepfakes' Trigger a Race to Fight Manipulated Photos and Videos."

18 Robitzski, "DARPA Spent $68 Million on Technology to Spot Deepfakes."

19 Knight, "A New Deepfake Detection Tool Should Keep World Leaders Safe—for Now."