Deepfakes are nearly seamless video and audio forgeries produced by artificial intelligence programs that yield seemingly realistic but fabricated images and sounds that portray people doing and saying things that never happened. Deepfake technology produces forgeries that are very difficult to detect, using software tools to create forgeries using existing images and videos that mimic the target’s facial expressions, movements, and the intonation, tone, stress, and rhythm of their speech in ways that are indistinguishable from real images and recordings.

Deepfakes gained public attention in 2017 when a Reddit user transposed celebrities’ faces onto actresses in pornographic videos. The technology is rapidly improving. Machine learning techniques are automating what used to be a manual process, allowing videos to be edited at machine-speed. It is much cheaper and simpler to make a deepfake now than it was two years ago, and deepfakes now require less data, time, and computational power. The technology is also much more accessible; open source software allows untrained persons to make rudimentary deepfakes easily. However, high-quality deepfakes still require professional expertise and specialized software.¹

**Trust and Screens**

Deepfakes are a new and powerful tool for falsehood. Their effect will depend not only on how persuasive they are but also on the timing and credibility of the repudiation of a deepfake. A compelling video released immediately before an election could do real harm before a denial could undo the damage.

People trust their senses too much. The internet does not provide—has never provided—a good way to distinguish between false and real. Simply looking at an image on a screen is not a good way to determine what is real and what is fake. Human senses are inadequate for detecting a well-tailored fraud. The ease of

---

anonymous actions online aggravates the problem, since not knowing the real source of a deepfake makes it harder to distinguish between real and false. Finally, the unmediated nature of publication online means that users can post material available to millions without any review to decide whether something is trustworthy or not.

Deepfakes are another part of the larger problem created by an untrustworthy medium. The internet provides immense value, but it is easily exploited for malicious purposes. Extremist groups and authoritarian opponents are among the beneficiaries of the internet’s “democratization” of knowledge and communication. Deepfakes give malicious actors a new tool for creating mistrust and confusion among the credulous or distrusting.

The Immediate Threat Is to Individuals

The immediate threat from deepfakes is to vulnerable individuals. A 2019 survey of 15,000 online deepfake videos found that 96 percent were pornographic. In one instance, unknown perpetrators created a pornographic deepfake of an Indian investigative journalist that went viral in an attempt to discredit her. Deepfakes will also enhance social engineering and phishing. There are already early instances of this. Criminals scammed a British energy company for $243,000 by spoofing a high-level executive's voice with AI and calling to demand payment.

Deepfakes expand the already dangerous ability to use the internet for fraud. For instance, in June 2018, eight people were killed in riots in India in reaction to a video circulating on WhatsApp about alleged child kidnappers. The video was from a Pakistani child safety campaign, but it had been framed as a child kidnapping in India. A picture of detained children provoked outrage against immigration policies until it became clear that it was from 2014. It doesn’t take deepfake technology to fool the public online.

The Threat to Democracies Is Growing

Deepfakes can be used as another tool to undermine democratic processes. The 2019 Worldwide Threat Assessment produced by the U.S. intelligence community warned that “adversaries and strategic competitors probably will attempt to use deepfakes or similar machine-learning technologies to create convincing—but false—image, audio, and video files to augment influence campaigns directed against the United States and our allies and partners.”

Deepfakes are a new form of an old problem. They are the latest in a long history of imagery and audio manipulation. Stalin was cutting political enemies out of photographs 80 years ago. In 1990, Adobe Photoshop gave people the ability to edit digital photos. That same year, Newsweek warned that authoritarian governments like China could get away with future atrocities like Tiananmen because, “with electronic photography they could deny the veracity of the newly malleable image.”

3. Rana Ayyub, “I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me,” HuffPost UK, November 21, 2018, https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316.
The internet allows these kinds of forgeries to propagate quickly and widely, and they build on techniques already used, such as doctored emails, selective editing or purloined documents, and leaks to mass social media audiences that target political opponents. The hyper-realism and rapid progress of deepfakes are a symptom of the larger problem of the growth of public distribution and the erosion of expertise in all forms of public dialogue, from fake news to rampant online conspiracy theories. One author even predicted, “we’re not so far from the collapse of reality.” A better way to think about this is that there can be multiple realities online, some intentionally falsified for political purposes or even simply mischief, and deepfakes expands the universe of online falsehood.

**Benign Uses**

Deepfakes have some benign uses, mainly in the entertainment industry. The most well-known deepfakes are probably the renderings of old characters from Star Wars films. It comes as no surprise, then, that Disney fought the state legislature of New York when they attempted to ban deepfakes. The technology also has medical applications as well, such as being used to generate new MRI images for medical training. Companies are exploring how to recreate the voices of deceased relatives. Lyrebird, a synthetic audio firm, has partnered with the ALS Association to help people with ALS regain control of their voice.

Other companies are exploring the commercial applications of the technology. Tencent, a Chinese technology firm, is experimenting with embedding new advertisements seamlessly into old movies; you could see a product placement for the next iPhone or Galaxy the next time you stream the first Terminator. Researchers are beginning to produce full body deepfakes which are already serving as synthetic news anchors in China.

**Defending Against Deepfakes**

The first step in defense lies with detection software and improving our technical ability to distinguish between fake and authentic video. Technical solutions include both AI systems trained to identify anomalies in a file that are characteristic of deepfakes, as well as cryptographic techniques that could be integrated into video and audio recording equipment to flag tampering. However, there is a limit to what technical solutions can accomplish. Methods to detect the current generation of deepfakes may become obsolete as the technology to produce deepfakes improves. Hany Farid, the image forensics expert who created PhotoDNA, explained, “we’re decades away from having forensic technology that [could] conclusively tell a real from a fake.” Finally, simply knowing that a certain video file has been manipulated does not tell us how to identify the creators and hold them accountable.

---

Better rules are also necessary, but there are no easy solutions. Virginia has passed a law criminalizing deepfakes used in revenge pornography, and Texas has criminalized deepfakes used to influence elections. Massachusetts and California considered but failed to pass bills aimed at countering deepfakes due to concerns about overregulation. At the federal level, bills have been introduced in both the House and Senate that would criminalize malicious creation and distribution of deepfakes, but the Senate did not pass them because of concerns about overreach and ineffective enforcement. The best outcome would be to develop technical and legal tools to manage the risk of harm from deepfakes without stifling innovation or free expression. Solutions should hold individual perpetrators accountable rather than instituting blanket bans, and we may need to consider—as with other false news and foreign propaganda—redefining the responsibilities of social media platforms. Some (like Reddit) have banned “involuntary pornography” and closed sites or lists dedicated to deepfakes, but these measures are reactive and not always comprehensive.

It is difficult to write laws that distinguish between malicious intent and parody and hard to identify deepfake creators and hold them accountable in court. Holding platforms that host deepfakes liable is unlikely to be successful until Section 230 of the Communications Decency Act, which grants immunity to platforms for most content they host, is amended. New laws need to focus not only on preventing deepfakes and punishing those who post them but also on addressing some of the core problems like authentication of identity—for trust online.

Arthur Nelson is program coordinator and research assistant with the Technology Policy Program at the Center for Strategic and International Studies (CSIS) in Washington, D.C. James Andrew Lewis is a senior vice president at CSIS.

This report is produced by the Center for Strategic and International Studies (CSIS), a private, tax-exempt institution focusing on international public policy issues. Its research is nonpartisan and nonproprietary. CSIS does not take specific policy positions. Accordingly, all views, positions, and conclusions expressed in this publication should be understood to be solely those of the author(s).

© 2019 by the Center for Strategic and International Studies. All rights reserved.

17. Shao, “Fake Videos Could Be the next Big Problem in the 2020 Elections.”