LAWFARE

PRIVACY: TECHNOLOGY

Deepfakes: A Looming Crisis for National Security, Democracy and Privacy?

By **Robert Chesney**, **Danielle Citron** Wednesday, February 21, 2018, 10:00 AM

Privacy Paradox: Rethinking Solitude

"We are truly fucked." That was Motherboard's spot-on reaction to deepfake sex videos (realistic-looking videos that swap a person's face into sex scenes actually involving other people). And that sleazy application is just the tip of the iceberg. As Julian Sanchez tweeted, "The prospect of any Internet rando being able to swap anyone's face into porn is incredibly creepy. But my first thought is that we have not even scratched the surface of how bad 'fake news' is going to get." Indeed.

Recent events amply demonstrate that false claims—even preposterous ones—can be peddled with unprecedented success today thanks to a combination of social media ubiquity and virality, cognitive biases, filter bubbles, and group polarization. The resulting harms are significant for individuals, businesses, and democracy. Belated recognition of the problem has spurred a variety of efforts to address this most recent illustration of truth decay, and at first blush there seems to be reason for optimism. Alas, the problem may soon take a significant turn for the worse thanks to deepfakes.

Get used to hearing that phrase. It refers to digital manipulation of sound, images, or video to impersonate someone or make it appear that a person did something—and to do so in a manner that is increasingly realistic, to the point that the unaided observer cannot detect the fake. Think of it as a destructive variation of the Turing test: imitation designed to mislead and deceive rather than to emulate and iterate.

I. Deepfakes and Harm to Individuals

Fueled by artificial intelligence, digital impersonation is on the rise. Machine-learning algorithms (often neural networks) combined with facial-mapping software enable the cheap and easy fabrication of content that hijacks one's identity—voice, face, body. Deepfake technology inserts individuals' faces into videos without their permission. The result is "believable videos of people doing and saying things they never did."

Not surprisingly, this concept has been quickly leveraged to sleazy ends. The latest craze is fake sex videos featuring celebrities like Gal Gadot and Emma Watson. Although the sex scenes look realistic, they are not consensual cyber porn. Conscripting individuals (more often women) into fake porn undermines their agency, reduces them to sexual objects, engenders feeling of embarrassment and shame, and inflicts reputational harm that can devastate careers (especially for everyday people). Regrettably, cyber stalkers are sure to use fake sex videos to torment victims.

What comes next? We can expect to see deepfakes used in other abusive, individually-targeted ways, such as undermining a rival's relationship with fake evidence of an affair or an enemy's career with fake evidence of a racist comment.

Blackmailers might use fake videos to extract money or confidential information from individuals who have reason to believe that disproving the videos would be hard (an abuse that will include sextortion but won't be limited to it). Reputations could be decimated, even if the videos are ultimately exposed as fakes; salacious harms will spread rapidly, technical rebuttals and corrections not so much.

All of this will be awful. A century ago, Justice Oliver Wendell Holmes warned of the danger of falsely shouting fire in a crowded theater. Now those false cries might go viral, fueled by the persuasive power of hyper-realistic evidence in conjunction with the distribution powers of social media.

But there's more to the problem than these individual harms. Deepfakes also have potential to cause harm on a much broader scale—including harms that will impact national security and the very fabric of our democracy.

II. Deepfakes as a Threat to National Security and Democracy

Deepfakes raise the stakes for the "fake news" phenomenon in dramatic fashion (quite literally). We have already seen trolls try to create panic over fake environmental disasters, and the recent Saudi-Qatar crisis may have been fueled by a hack in which someone injected fake stories (with fake quotes by Qatar's emir) into a Qatari news site. Now, let's throw in realistic-looking videos and audio clips to bolster the lies. Consider these terrifying possibilities:

- Fake videos could feature public officials taking bribes, uttering racial epithets, or engaging in adultery.
- Politicians and other government officials could appear in locations where they were not, saying or doing horrific things that they did not.
- Fake videos could place them in meetings with spies or criminals, launching public outrage, criminal investigations, or both.
- Soldiers could be shown murdering innocent civilians in a war zone, precipitating waves of violence and even strategic harms to a war effort.
- A deepfake might falsely depict a white police officer shooting an unarmed black man while shouting racial epithets.
- A fake audio clip might "reveal" criminal behavior by a candidate on the eve of an election.
- A fake video might portray an Israeli official doing or saying something so inflammatory as to cause riots in neighboring countries, potentially disrupting diplomatic ties or even motivating a wave of violence.
- False audio might convincingly depict U.S. officials privately "admitting" a plan to commit this or that outrage overseas, exquisitely timed to disrupt an important diplomatic initiative.
- A fake video might depict emergency officials "announcing" an impending missile strike on Los Angeles or an emergent pandemic in New York, provoking panic and worse.

Note that these examples all emphasize how a well-executed and well-timed deepfake might generate significant harm in a particular instance, whether the damage is to physical property and life in the wake of social unrest or panic or to the integrity of an election. The threat posed by deepfakes, however, also has a long-term, systemic dimension.

The spread of deepfakes will threaten to erode the trust necessary for democracy to function effectively, for two reasons. First, and most obviously, the marketplace of ideas will be injected with a particularly-dangerous form of falsehood. Second, and more subtly, the public may become more willing to disbelieve true but uncomfortable facts. Cognitive biases already encourage resistance to such facts, but awareness of ubiquitous deepfakes may enhance that tendency, providing a ready excuse to disregard unwelcome evidence. At a minimum, as fake videos become widespread, the public may have difficulty believing what their eyes (or ears) are telling them—even when the information is quite real.

In a recent report, the Belfer Center highlighted the national security implications of sophisticated forgeries. For instance, an adversary could acquire real (and sensitive) documents through cyber-espionage and leak the real documents along with forgeries supported by "leaked" forged audio and video. In similar fashion, public trust may be shaken, no matter how credible the government's rebuttal of the fake videos. Making matters worse, news organizations may be chilled from rapidly reporting real, disturbing events for fear that the evidence of them will turn out to be fake (and one can well-imagine someone trying to trap a news organization in exactly this way).

III. Who Will Cause These Harms?

The capacity to generate hyper-realistic deepfakes will open up remarkable opportunities for covert action, plainly. Hostile foreign intelligence agencies already will be quite aware of this (as will our own intelligence agencies, no doubt), and if 2016 taught us nothing else it should at least have taught us to expect that at least some of those foreign intelligence agencies will make sophisticated efforts to exploit such possibilities.

But there is no reason at all to think that the capacity to generate persuasive deepfakes would stay with governments. Indeed, there's no reason to think the capacity even will originate with governments. Private sector and academic researchers may be first past the post. Consider the recent example of University of Washington researchers who created a tool that alters videos so speakers say something different from what they said at the time. Researchers used a video of former president Barack Obama speaking to show how the tool can be used effectively thanks to the existence of plentiful video footage needed for training the algorithms. Interested readers can find the study here.

At any rate, the capacity to generate persuasive deepfakes (and, critically, user-friendly software enabling almost anyone to exploit that capacity) will diffuse rapidly and globally (in keeping with the dynamics Benjamin Wittes and Gabriella Blum explore in their compelling book *The Future of Violence*). Thus, even if it does not start that way, the technology will end up in the hands of a vast range of actors willing to use deepfakes in harmful ways.

IV. Responding to the Threat: The Limits of Technological and Legal Solutions

The ideal response to the deepfake threat would be the simultaneous development and diffusion of software capable of rapidly and reliably flagging deepfakes, and then keeping pace with innovations in deepfake technology. If such technology exists and is deployed through the major social media platforms especially, it would go some way towards ameliorating the large-scale harms described above (though it might do little to protect individuals from deepfake abuses that don't require distribution-at-scale through a gatekeeping social media platform).

Unfortunately, it is not clear that the defense is keeping pace for now. An arms race to fortify the technology is on, but Dartmouth professor Hany Farid, the pioneer of PhotoDNA (a technology that identifies and blocks child pornography), warns: "We're decades away from having forensic technology that … [could] conclusively tell a real from a fake. If you really want to fool the system you will start building into the deepfake ways to break the forensic system." This suggests the need for an increase—perhaps a vast increase—in the resources being devoted to the development of such technologies.

What about civil and criminal law? To be sure, law might have *some* deterrent and compensatory role, at least as to domestic actors and others who might be brought within reach of U.S. legal process. Defamation law, for instance, provides redress for knowing or reckless falsehoods that damage reputation of public officials or public figures, and criminal law in various ways could be brought to bear. Law's deterrent value, however, would be modest at best because as Benjamin Wittes and Danielle Citron have explored, Section 230 of the Communications Decency Act immunizes from (most) liability the entities best situated to minimize damage efficiently: the platforms.

But just like we might wonder what deterrent effect will really follow from Robert Mueller's recent indictment of Russians who interfered in the 2016 election, we might similarly wonder about the prospects for using law to deter or punish the larger-scale deepfake harms described above when they flow from foreign sources. Is there anything else that might work?

V. The Promise and Peril of Immutable Authentication Trails

Consider the worst-case scenario: We enter a world in which it becomes child's play to portray people as having done or said things they did not say or do; we lack the technology to reliably expose the fakes; and we lack the legal and practical capacity to punish and deter use of deepfakes to inflict individual and large-scale harms. In that case, it is not hard to imagine the rise of a profitable new service: immutable authentication trails.

The idea is this: A person who is sufficiently interested in protecting against a targeted deepfake (or whose employer feels this way) may prove willing to pay for a service that comprehensively tracks some or all of the following—their movements, electronic communications, in-person communications, and surrounding visual circumstances. The vendor providing the service, to be successful, would have to develop a sufficient reputation for the immutability and comprehensiveness of its data. It might then have its own arrangements with media platforms allowing it to debunk—perhaps quite rapidly—emergent deep fakes impacting its clients. If successful, it is not hard to imagine it proving somewhat popular (especially with employers who might require assent to such a service as a term of employment, barring legal obstacles to doing so).

Whatever the benefits, the social cost should such a service emerge and prove popular would be profound. It risks the "unraveling of privacy"—that is, the collapse of privacy by social consent regardless of what legal protections for privacy there may be.

As an initial matter, the vendor supplying such a service and maintaining the resulting data would be in an extraordinary position of power both as to its individual clients (obviously) and as both a market force and a research force (unless it eschews any effort to exploit what could well became a database of human behavior of unprecedented depth and breadth, or as Paul Ohm presciently put it, a "database of ruin").

What's more, the third-party doctrine (at least as things currently stand) presumably would make that database relatively accessible to government's investigative authorities—and those authorities would not be very good at their jobs if they did not begin to look, first and foremost, towards this rich trove of information whenever investigating someone. And while it is possible that doctrinal developments at the Supreme Court (in *Carpenter* and its eventual progeny) may yet require the government to seek warrants to obtain such comprehensive data-collections held by third parties, or that Congress might do the same via statute as applied to such a new and impactful service industry, the fact remains that—once the right legal process is used—the government's capacity to know all about a suspect would be unrivaled as a historical matter (especially as combined with other existing aggregations of data). That in turn will often produce remarkable opportunities to learn of *other* misdeeds by a suspect; other information that might not be inculpatory as such but that might nonetheless provide important leverage over the suspect; and (bearing in mind that much of the data will involve interactions with other people) similar information about other persons apart from the suspect.

Perhaps such a system would yield more good than harm on the whole and over time (particularly if there is legislation well-tailored to regulate access to such a new state of affairs). Perhaps time will tell. For now, our aim is no more and no less than to identify the possibility that the rise of deepfakes will in turn give birth to such a service, and to flag the implications this will have for privacy. Enterprising businesses may seek to meet the pressing demand to counter deepfakes in this way, but it does not follow that society should welcome—or wholly accept—that development. Careful reflection is essential now, before *either* deepfakes *or* responsive services get too far ahead of us.

Topics: Privacy: Technology

Tags: deep fakes, deepfake

Bobby Chesney is the **Charles I. Francis Professor in Law** and Associate Dean for Academic Affairs at the University of Texas School of Law. He also serves as the Director of UT-Austin's interdisciplinary research center the Robert S. Strauss Center for International Security and Law. His scholarship encompasses a wide range of issues relating to national security and the law, including detention, targeting, prosecution, covert action, and the state secrets privilege; most of it is posted **here**. Along with Ben Wittes and Jack Goldsmith, he is one of the co-founders of the blog.

@bobbychesney

Danielle Citron is a Professor of Law at Boston University School of Law and a 2019 MacArthur Fellow. She is the author of *Hate Crimes in Cyberspace* (Harvard University Press 2014).