Don’t Believe It If You See It: Deep Fakes and Distrust

Author : Kristen Eichensehr

Date : September 27, 2018

Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 **Cal. L. Rev.** __ (forthcoming 2019), available at SSRN.

It’s no secret that the United States and much of the rest of the world are struggling with information and security. The flow of headlines about data breaches, election interference, and misuse of Facebook data show different facets of the problem. Information security professionals often speak in terms of the “CIA Triad”: confidentiality, integrity, and availability. Many recent cybersecurity incidents involve problems of confidentiality, like intellectual property theft or theft of personally identifiable information, or of availability, like distributed denial of service attacks. Many fewer incidents (so far) involve integrity problems—instances in which there is unauthorized alteration of data. One significant example is the Stuxnet attack on Iranian nuclear centrifuges. The attack made some centrifuges spin out of control, but it also involved an integrity problem: the malware reported to the Iranian operators that all was functioning normally, even when it was not. The attack on the integrity of the monitoring systems caused paranoia and a loss of trust in the entire system. That loss of trust is characteristic of integrity attacks and a large part of what makes them so pernicious.

Bobby Chesney and Danielle Citron have posted a masterful foundational piece on a new species of integrity problem that has the potential to take such problems mainstream and, in the process, do great damage to trust in reality itself. In *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, Chesney and Citron explain a range of possible uses for “deep fakes,” a term that originated from imposing celebrities’ faces into porn videos, but that they use to describe “the full range of hyper-realistic digital falsification of images, video, and audio.” (P. 4.)

After explaining the technology that enables the creation of deep fakes, Chesney and Citron spin out a parade of (plausible) horribles resulting from deep fakes. Individual harms could include exploitation and sabotage, such as a fake compromising video of a top draft pick just before a draft. (P. 19.) The equally, if not more, worrisome societal harms from deep fakes include manipulating elections through timely release of damaging videos of a candidate, eroding trust in institutions though compromising videos of their leaders, exacerbating social divisions by releasing videos of police using racial slurs, spurring a public panic with recordings of government officials discussing non-existent disease outbreaks, and jeopardizing national security through videos of U.S. troops perpetrating atrocities. (Pp. 22-27.)

So what can be done? The short answer appears to be not much. The authors conclude that technology for detecting deep fakes won’t save us, or at least won’t save us fast enough. Instead, they “predict,” but don’t necessarily endorse, “the development of a profitable new service: immutable life logs or authentication trails that make it possible for the victim of a deep fake to produce a certified alibi credibly proving that he or she did not do or say the thing depicted.” (P. 54.) This possible “fix” to the problem of deep fakes bears more than a passing resemblance to the idea of “going clear” spun out in Dave Eggers’ book *The Circle*. (Pp. 239-42.) In the novel, politicians begin wearing 24-hour electronic monitoring and streaming devices to build the public’s trust—and then others are pressured to do the same because, as Eggers puts it, “If you aren’t transparent, what are you hiding?” (P. 241.) When the “cure” for our problems comes from dystopian fiction, one has to wonder whether it’s worse than the

disease. Moreover, companies offering total life logs would themselves become ripe targets for hacking (including attacks on confidentiality and integrity) given the tremendous value of the totalizing information they would store.

If tech isn't the answer, what about law? Chesney and Citron are not optimistic about most legal remedies either. They are pessimistic about the ability of federal agencies, like the Federal Trade Commission or Federal Communications Commission, to regulate our way out of the problem. They do identify ways that criminal and civil remedies may be of some help. Victims could sue deep fake creators for torts like defamation and intentional infliction of emotional distress, and deep fake creators might be criminally prosecuted for things like cyberstalking (18 U.S.C. § 2261A) or impersonation crimes under state law. But, as the authors note, legal redress even under such statutes may be hampered by, for example, the inability to identify deep fake creators, or to gain jurisdiction over them. These statutes also do little do redress the societal, as opposed to individualized, harms from deep fakes.

For deep fakes perpetrated by foreign states or other hostile actors, Chesney and Citron are somewhat more optimistic, highlighting the possibility of military and covert actions, for example, to degrade or destroy the capacity of such actors to produce deep fakes. (Pp. 49-50.) They also suggest a way to ensure that economic sanctions are available for "attempts by foreign entities to inject false information into America's political dialogue," including attempts using deep fakes. (P. 53.) These tactics might have some benefit in the short term, but sanctions have not yet stemmed efforts at foreign interference in elections. And efforts to disrupt Islamic State propaganda have shown that attempts at digital disruption of adversaries' capacities may often prompt a long-running battle of digital whack-a-mole.

One of the paper's most interesting points is its discussion of another tactic that one might think would help address the deep fake problem, namely, public education. Public education is often understood to help inoculate against cybersecurity problems. For example, teaching people to use complex passwords and not to click on suspicious email attachments bolsters cybersecurity. But Chesney and Citron point out a perverse consequence of educating the public about deep fakes. They call it the "liar's dividend": "a skeptical public will be primed to doubt the authenticity of real audio and video evidence," so those caught engaging in bad acts in authentic audio and video recordings will exploit this skepticism to "try to escape accountability for their actions by denouncing authentic video and audio as deep fakes." (P. 28.)

Although the paper is mostly profoundly disturbing, Chesney and Citron try to end on a positive note by focusing on the content screening and removal policies of platforms like Facebook. They argue that the companies' terms of service agreements "will be primary battlegrounds in the fight to minimize the harms that deep fakes may cause," (P. 56) and urge the platforms to practice "technological due process." (P. 57.) Facebook, they note, "has stated that it will begin tracking fake videos." (P. 58.) The ending note of optimism is welcome, but rather underexplored in the current draft, leaving readers hoping for more details on what, when, and how much the platforms might be able and willing to do to prevent the many problems the authors highlight. It also raises fundamental questions about the role of private companies in playing at least arguably public functions. Why should this be the companies' problem to fix? And if the answer is because they're the only ones who can, then more basically, how did we come to the point where that is the case, and is that an acceptable place to be?

In writing the first extended legal treatment of deep fakes, Chesney and Citron understandably don't purport to solve every problem they identify. But in a world plagued by failures of imagination that leave the United States reeling from unexpected attacks—Russian election interference being the most salient—there is tremendous benefit to thoughtful diagnosis of the problems deep fakes will cause. Deep fakes are, as Chesney and Citron's title suggests, a "looming challenge" in search of solutions.

Cite as: Kristen Eichensehr, *Don't Believe It If You See It: Deep Fakes and Distrust*, JOTWELL (September 27, 2018) (reviewing Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 **Cal. L. Rev.** __ (forthcoming 2019), available at SSRN), https://cyber.jotwell.com/dont-believe-it-if-you-see-it-deep-fakes-and-distrust/.