2019

HOOVER INSTITUTION

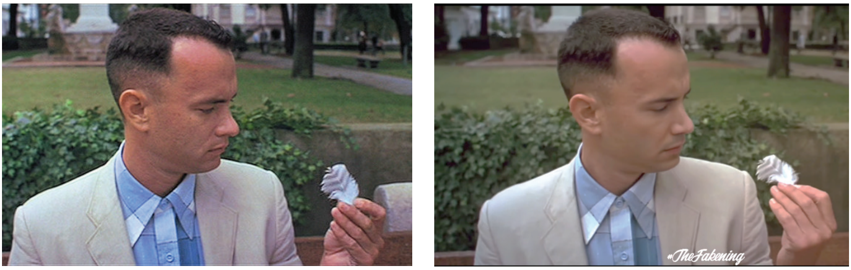
Summer Policy Boot Camp

Director's Award Winners

Less Can Be More When Regulating Deepfakes

By Timothy Anderson, Department of Electrical Engineering, Stanford University

On November 2, 2017, a forum on Reddit.com was created for posting and discussing pornographic videos known as "deepfakes" which used AI to map faces of female celebrities such as Gal Gadot and Maisie Williams onto pre-existing video clips. The Reddit page quickly exploded to several thousand visitors—many of whom created and posted new videos—and gained wide attention after being featured in *Vice* Motherboard.[1] While the Reddit page was shut down after only three months, the deepfakes phenomenon laid bare to the broader public the superhuman capabilities of readily-available AI, the ethical bankruptcy of modern tech culture, and the chilling reality that anyone with basic programming skills and access to a modern computer now has the power to violate innocent others or disrupt national politics without leaving their desk.



Left: Original image from *Forrest Gump*.
Right: Deepfake video of Keanu Reeves mapped onto Tom Hanks.[2]

Computers Learn to Trick You

Deepfakes are built on "deep neural networks," a broad family of AI algorithms that are nearly ubiquitous in today's data processing systems, from search engine results to voice assistants and automated photo tagging. The neural networks behind deepfakes differ from other AI in that, instead of classifying data such as words in an audio clip, these networks, known as "generative networks," learn to generate new data that is indistinguishable from existing real-world examples.[3] Generative network–based algorithms have become some of the most powerful in AI and are now capable of creating believable fake data in a wide range of areas.

An interesting application of these algorithms is translating images between domains, such as daytime to nighttime photos or Monet's style to Van Gogh's. However, as virtual reality pioneer Young Harvill said about his invention, "a lot of the killer app for some of the first technologies is porn,"[4] and unfortunately, image translation proved no exception. The same technology that can change horses to look like zebras can also change the face in a photograph; less than a year after publication of the hallmark papers on image translation with generative networks,[5] the first deepfakes were posted to Reddit.

Artificial Intelligence for Social Ill

If an anonymous Reddit user can create realistic compromising footage of celebrities, what would stop a vengeful ex-lover from creating footage of you? What about a plaintiff submitting AI-fabricated evidence to alter a court case? A rogue actor creating a fake video of the president declaring war?

Deepfakes have been justifiably denounced as everything from misogynistic and exploitative toward women[6] to a national security disaster in the making,[7] as well as an existential threat to the idea of video evidence and news itself.[8] The growth in nefarious applications of AI is largely because AI algorithms are now so easy to build. Today, thanks to free software maintained by Google, Facebook, and Nvidia, anyone with a college freshman's knowledge of mathematics and computer programming can create their own AI algorithms. Such resources have been a boon for research in AI and other data-centric fields, but the reduced barriers to entry also mean that bad actors everywhere now have a new tool at their disposal.

Existing Laws Can Already Regulate Deepfakes

The public has begun to demand action from lawmakers on deepfakes, especially in the wake of viral fake videos of Mark Zuckerberg and Nancy Pelosi,[9] and Congress has started to take action. In December 2018, Ben Sasse (R-NE) introduced the Malicious Deep Fake Prohibition Act, which would criminalize the creation of certain types of deepfakes.[10] House Foreign Intelligence Committee chairman Adam Schiff has also begun holding congressional hearings related to deepfakes and other disruptive AI-based technologies.[11]

However, as John Villasenor points out, we likely already have legal frameworks to regulate and prosecute many problematic types of deepfakes.[12] While generative networks are often hyped as having "the gift of imagination,"[13] this is not actually the case. The single most important trait of today's AI is that it cannot invent: AI can only imitate the real-world data that it ingests. While subtle, this aspect of AI directly leads to how we can define, regulate, and—if necessary—prosecute malicious deepfakes.

One domain of concern is false advertising. Given the availability of deepfake audio generators for some popular internet personalities such as Joe Rogan, it would be trivial for a corporation to synthesize and publish a fake endorsement from one of these hugely influential individuals. However, false advertising is already forbidden by the Lanham Act: "Any person who . . . uses in commerce any word, term, name, symbol, or device . . . false or misleading representation of fact . . . shall be liable in a civil action."[14] The endorser's image or voice was inherently used in training the deepfake algorithm, and as such, the creator of a deepfake endorsement could be held civilly liable for misuse under existing law.

Perhaps the most morally abhorrent application of deepfakes is hyperrealistic simulated child pornography. Again though, this application would require the use of real images of children at some point in building the generative network algorithm, and therefore would be prosecutable under the PROTECT Act of 2003, which prohibits the use of images of minors in the creation of simulated child pornography.

Advancing Broader AI Efforts Can Also Fight Deepfakes

The uncomfortable reality is that the best course of action for policymakers may be to do very little to address deepfakes directly. Existing laws cover many harmful applications of deepfakes, and banning certain deepfakes would only hamper research into benevolent technologies employing similar algorithms. The threat from deepfakes is simply the worst-case end result of the democratization and proliferation of AI. The cultural and technological environment that created deepfakes—not the resultant algorithms themselves—should be the focus of any policy. To this end, there are four key areas in which we can concentrate our efforts to combat deepfakes:

Incentivize industry to improve deepfake-detection algorithms. Incentivizing industry to prevent deepfakes is the one arena in which direct action may be the best course. All malicious deepfakes falling under US law will be published on privately owned internet platforms, so detection and removal of deepfakes must necessarily be done by technology companies. We must require private technology companies to implement deepfake safeguards on their platforms, and fine those who fail to remove deepfakes violating existing laws. While some companies such as Reddit and Twitter already have voluntarily banned malicious deepfakes, others such as Facebook and internet pornography giant MindGeek have yet to take significant action. Fining these companies would provide a monetary incentive for preventing deepfakes.

Implementing a policy of this sort would require the government to define and regularly update what is considered to be state-of-the-art deepfakes generation and detection techniques. The Deepfakes Reports Act, introduced in July 2019 by US Senator Gary Peters (D-MI),[15] is a huge step towards developing such a standard. This law would require the Department of Homeland Security to produce regular reports on deepfakes, which in turn could be used to delineate what illicit content should be detectable by private companies' filters.

Expand fundamental AI research at government and academic institutions. While we must significantly expand government-backed research specifically into deepfakes,[16] this should be part of a much broader effort to improve American AI capabilities. With Russia and China making large-scale government-backed efforts into AI research, maintaining AI parity is now a national security concern.

However, the unfortunate reality of AI in the United States is that most of the intellectual and computational resources are held by private industry. If we plan to keep up in the global AI arms race, the United States must greatly expand research into AI relevant to national security. Creating programs such as DARPA's Media Forensics Office, which focuses on research related to fabricated images and video, is a step in the right direction.[17] Beyond expanding research at government laboratories, we should also increase research grants from agencies such as the National Science Foundation for projects focusing on national security–relevant AI.

Improve AI capabilities by improving education. Alongside our research capabilities, we must expand our domestic intellectual capital in AI. While the United States is home to many of the best AI research groups and technical universities, AI is dominated by non-Americans working or studying at American universities and corporations. While this is certainly no issue for the advancement of AI as a field, it is a severe concern for America's ability to integrate AI into our military and intelligence operations due to citizenship requirements for working in these areas. To maintain AI parity with other great powers, the United States must grow the pool of American students entering college with sufficient math and computer science backgrounds to succeed in studying AI.

The inability of American undergraduates to meet the demand for AI engineering jobs traces back to poor secondary education, particularly in STEM fields. However, improving domestic AI intellectual capital and readiness need not involve a comprehensive overhaul of the education system. To address this challenge, it could be sufficient to simply fund summer programs for a few thousand students to learn programming and quantitative skills relevant to AI. Such programs could serve the dual purpose of preparing top students to be competitive in technical fields at the university level while providing a prestigious summer opportunity for low-income high school students who would otherwise be unable to afford the often expensive summer intensives offered at many top universities.

Promote ethics education alongside technical fields. Finally, the threat of deepfakes highlights the need to include law and ethics education alongside technical instruction. All unethical technology begins with an unethical engineer: instead of simply banning unethical AI, we should address the dearth of ethical concern shown by many of today's inventors and entrepreneurs. One approach would be to provide curriculum development grants to universities who integrate ethics education into their introductory computer science curricula. By exposing students to the concerns and perils surrounding technology early in their careers, we could help to embed technology ethics in American tech culture.

Cynics may argue that expecting self-regulation of the tech industry is foolhardy—the prevailing ethos of Silicon Valley is Facebook's unofficial motto "Move fast and break things"[18]—but recent events show that bringing AI ethics into the public conscience is having an effect. For example, when DeepNude, a program that digitally undressed women, was published earlier this year, GitHub, the largest website for publishing computer code and now a Microsoft subsidiary, swiftly banned this app and any related project.[19] There still exists much work to be done in this area, but tech culture in the United States is slowly shifting in a promising direction.

Summary and Conclusion

While deepfakes indeed pose a threat to our privacy and security, this threat is far from insurmountable. When taking action on deepfakes, our legislators should eschew rash solutions that will be detrimental to long-term AI advances and instead pursue policies that address the underlying issues in technology culture and education. By incentivizing private industry against deepfakes, promoting AI research, and improving equity in and the ethics components of domestic STEM education, we can effectively mitigate deepfakes, strengthen our domestic AI capabilities, and continue to foster the environment of innovation which gave us such frightening and fascinating technology in the first place.

[1] Samantha Cole , "AI-Assisted Fake Porn Is Here and We're All F——d," Vice, December 11, 2017, https://www.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn.

[2] "Keanu Reeves as Forest [sic] Gump Deepfake—It's Breathtaking!" YouTube, https://www.youtube.com/watch?v=cVljNVV5VPw.

[3] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, and Aaron Courville, "Generative Adversarial Networks," 2014, arXiv:1406.2661.

[4] Adam Fisher, Valley of Genius: The Uncensored History of Silicon Valley (As Told by the Hackers, Founders, and Freaks Who Made It Boom (New York: Twelve, 2018).

[5] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," 2016, arXiv:1611.07004; Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," 2017, arXiv:1703.10593.

[6] Drew Harwell, "Fake-porn videos are being weaponized to harass and humiliate women: 'Everybody is a potential target,'" Washington Post, December 30, 2018, https://www.washingtonpost.com/technology/2018/12/30/fake-porn-videos-are-being-weaponized-harass-humiliate-women-everybody-is-potential-target/?noredirect=on.

[7] James Vincent, "Watch Jordan Peele use AI to make Barack Obama deliver a PSA about fake news," *The Verge*, April 17, 2018, https://www.theverge.com/tldr/2018/4/17/17247334/ai-fake-news-video-barack-obama-jordan-peele-buzzfeed.

[8] Riana Pfefferkorn, "“Deepfakes': A New Challenge for Trial Courts," *NWSidebar*, June 24, 2019, https://nwsidebar.wsba.org/2019/03/13/deepfakes-a-new-challenge-for-trial-courts; Oscar Schwartz, "You thought fake news was bad? Deep fakes are where truth goes to die," *The Guardian*, November 12, 2018, https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth.

[9] Robert Chesney, Danielle Citron, and Quinta Jurecic, "About That Pelosi Video: What to Do About 'Cheapfakes' in 2020," *Lawfare*, June 2, 2019, https://www.lawfareblog.com/about-pelosi-video-what-do-about-cheapfakes-2020.

[10] Malicious Deep Fake Prohibition Act of 2018, S.3805, introduced December 21, 2018, 115th Congress (2017–2018), https://www.congress.gov/bill/115th-congress/senate-bill/3805.

[11] Elizabeth Culliford, "House Intelligence chief presses social media companies on deepfake policies," *Reuters*, July 15, 2019, https://www.reuters.com/article/us-usa-election-deepfakes/house-intelligence-chief-presses-social-media-companies-on-deepfake-policies-idUSKCN1UA2GC; Eric Johnson, "“We're not ready' for foreign election interference in 2020, says Rep. Adam Schiff," *Vox*, July 22, 2019, https://www.vox.com/recode/2019/7/22/20702196/adam-schiff-deepfakes-nancy-pelosi-google-twitter-facebook-2020-youtube-kara-swisher-decode-podcast.

[12] John Villasenor, "Artificial intelligence, deepfakes, and the uncertain future of truth," Brookings Institution, February 14, 2019, https://www.brookings.edu/blog/techtank/2019/02/14/artificial-intelligence-deepfakes-and-the-uncertain-future-of-truth.

[13] Martin Giles, "The GANfather: The man who's given machines the gift of imagination," *MIT Technology Review*, February 27, 2018, https://www.technologyreview.com/s/610253/the-ganfather-the-man-whos-given-machines-the-gift-of-imagination.

[14] 15 U.S.C. § 1125.

[15] "Peters, Colleagues Introduce Bipartisan Bill to Respond to Growing Threat of Deepfakes," Homeland Security & Governmental Affairs, Minority Media, July 3, 2019, https://www.hsgac.senate.gov/media/minority-media/peters-colleagues-introduce-bipartisan-bill-to-respond-to-growing-threat-of-deepfakes.

[16] Villasenor, "Artificial Intelligence."

[17] Matt Turek, "Media Forensics (MediFor), Defense Advanced Research Projects Agency (DARPA), https://www.darpa.mil/program/media-forensics.

[18] Fisher, *Valley of Genius*.

[19] Arwa Mahdawi, "An app using AI to 'undress' women offers a terrifying glimpse into the future, *The Guardian*, June 29, 2019, https://www.theguardian.com/commentisfree/2019/jun/29/deepnude-app-week-in-patriarchy-women; Joseph Cox, "GitHub Removed Open Source Versions of DeepNude," *Vice* Motherboard, July 9, 2019, https://www.vice.com/en_us/article/8xzjpk/github-removed-open-source-versions-of-deepnude-app-deepfakes.