Online Ethics Center
FOR ENGINEERING AND SCIENCE

# Deepfakes and the Value-Neutrality Thesis

## Author(s)

Michael J. Quinn
Nathan Colaner

## Description

The creator of FakeApp says it would be wrong to condemn the technology. How could we not?

## Body

A “deepfake” is a video or still image of a person that is modified to depict someone else. The term comes from the alias of Reddit user “deepfakes,” who in late 2017 used open-source machine learning tools to put the faces of Scarlett Johansson, Maisie Williams, and other celebrities on the bodies of women in pornographic videos and then posted the videos on Reddit [1]. It didn’t take long before another Reddit user, “deepfakeapp,” published FakeApp, an application making it possible for less-tech-savvy computer users to create their own deepfakes [2].

Samantha Cole wrote several articles that brought attention to the pornographic deepfakes circulating online, and in February 2018 Reddit banned the community of users sharing deepfake videos because they violated Reddit’s policy against the posting of nonconsensual pornography. Twitter and Pornhub followed suit [3].

# Social Harms

Although researchers are striving to create tools to detect deepfakes, the harms will likely be merely slowed, rather than stopped [4]. The personal harms that can be unleashed by deepfakes are limited only by the imaginations of bad actors, but they are dwarfed by the scale of societal harms we may soon experience. An obvious misuse of deepfakes is their potential role in creating more sophisticated fabricated news stories. Conversely, another harm of deepfakes is that they can cast doubt on authentic stories [4]. It is chilling to think of the effect they may have on the 2020 US presidential election.

Even if deepfakes are identified and pointed out, that may not matter. On January 6, 2020, Representative Paul Gosar of Arizona tweeted a deepfake photo of Barack Obama with Iranian president Hassan Rouhani with the caption, “The world is a better place without these guys in power,” presumably as a justification of the killing of Iranian General Soleimani. When called out for disseminating a deepfake, Gosar protested that he never *actually* claimed that the photo was real [5].

Could such blatant gaslighting be effective? Yes. Research shows that political campaigns are not effective at persuading voters to change their views. Instead, what campaigns *can* do is reinforce voters’ political orientations and motivate them to vote [6]. Despite the fact that President Obama never met Iranian President Rouhani and President Rouhani is still in power, Gosar’s tweet may have had its intended effect. This is in line with another recent data point: the 2016 US presidential election demonstrated that people are drawn to, and share, stories that confirm their political views, even if those stories are false.

# A New Ethical Challenge

There are several distinct legal and ethical challenges posed by the revolution brought on by artificial intelligence. Some of the challenges are futuristic, but the advent of deepfake technology forces us to face an imminent one: trust in what is real. A reasonable fear is that the overall effect of deepfakes will be to undermine a “shared sense of reality” that is essential to a healthy democracy [7]. But it is important to note the sense in which this is a genuinely new problem.

For as long as the prefrontal cortex has been around, there have been different legitimate ways to interpret the events around us. There are an indefinite number of reasons for this, and the challenge of conflicting interpretations will never stop. But at least our social structures and institutions, and even biological evolutionary development itself, have already accounted for this level of disagreement. And so different people interpreting reality differently is nothing new.

Likewise, the problem of false accounts is not a new problem. “Fake news,” for example, does indeed go one step further than the problem of conflicting interpretation of reality; it represents attempts to create conflicting narratives. The influence of “fake news” is rightfully disconcerting, but for all that, fake stories have always been with us, and so their influence is not new.

Deepfakes, however, go further. The deception is not simply in the form of false narratives; these new deceptions have a visceral quality. This is because deepfakes are “not just lies, but ones that betray sight and sound, two of our most innate and cherished senses” [8, p. 961]. The problem of a *visceral* unshared reality, as opposed to an unshared reality created by false narratives, is a new, unaccounted-for development that we are going to have to scramble to address. The human brain, with all of its evolved sophistication, does not have an obvious way to dismiss AI-generated sights and sounds when they are every bit as enticing as organically generated sights and sounds. We are going to need new laws, new cultural norms, and new standards for belief. None of this promises to be easy.

# Technology and Value-Neutrality

The creator of FakeApp said, “I’ve given it a lot of thought, and ultimately I’ve decided I don’t think it’s right to condemn the technology itself – which can of course be used for many purposes, good and bad” [9]. That’s a classic line: “Technology is value-neutral.” For example, the book *Irresistible* by sociologist Adam Alter focuses on behavioral addiction brought about by our engagements with screens. In this context, he claims that “[t]ech isn’t morally good or bad until it’s wielded by the companies that fashion it for mass consumption” [10, p. 8]. The general point is a familiar one: “tech is not inherently good or bad” [10, p. 316].

This position can mean two things. The more robust way to interpret it is as a claim about technology in general – in other words, *any* technology whatsoever is value-

neutral. A more limited interpretation is that it merely claims that *certain* technologies can be value-neutral, even while others are not.

The claim that any possible technology whatsoever is value-neutral is more difficult to accept, morally speaking. It would serve as an excuse for the designers of any technology, such that they would have no responsibility whatsoever to account for the ways that their creations could be misused. A vital imperative of our AI future, perhaps the one most likely to keep civilization intact – is that the developers of AI technology design technology with constant concern to how that technology could be misused or generate unintended consequences.

# Evaluating Deepfake Technology

What, then, about this technology in particular? Is the technology that creates deepfakes value-netural? In order to make the case that a *particular* technology is value neutral, at the very least its potential benefits should be comparable to its potential harms. Remember that the FakeApp creator excused himself with the line that it “can of course be used for many purposes, good and bad.” In the absence of further evidence of beneficial consequences, we should conclude that this line is obviously false.

The potential benefits of deepfakes, primarily in entertainment, are dwarfed by their potential harms. Ultimately, deepfakes are about deceit. The goal of the effort is to create fantasies that cannot be distinguished from reality. Putting words into someone else’s mouth to influence an election or inserting someone into a pornographic movie are morally reprehensible actions. The creators of the tools used for these purposes – FakeApp, Faceswap, and DeepFaceLab – are morally responsible for the harms deepfakes have caused and are likely to cause to innocent people and to our democratic institutions.

Assuming that it is unrealistic to completely halt the development of such technologies, a reasonable first step is strict, responsible regulation. But perhaps more importantly, like all technology, the designers must abandon the value-neutrality thesis and anticipate the cases *when* – not *if* – their technology will be used for evil.

[1] Samantha Cole. “AI-Assisted Fake Porn Is Here and We’re All Fucked.” *Motherboard: Tech by Vice.* December 11, 2017. vice.com.

[2] Samantha Cole. “We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now.” Motherboard: Tech by Vice. January 24, 2018. vice.com.

[3] Erin Carson. “Reddit Cracks Down on ‘Deepfake’ Pornography.” CNET. February 7, 2018. cnet.com.

[4] Cade Metz. “Internet Companies Prepare to Fight the ‘Deepfake’ Future.” *The New York Times*. November 24, 2019. nytimes.com.

[5] Linda Qiu. “Republican Congressman Shares Fake Image of Obama and Iranian President.” *The New York Times*. January 6, 2020. nytimes.com.

[6] Joshua L. Kalla and David E. Broockman. “The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments.” *American Political Science Review* 112, 1. 2018.

[7] Anamitra Deb, Stacy Donohue, and Tom Glaisyer. “Is Social Media a Threat to Democracy?” The Omidyar Group. October 1, 2017.

[8] Jessica Silbey and Woodrow Hartzog. “The Upside of Deep Fakes.” *Maryland Law Review* 78, 4. 2019.

[9] Kevin Roose. “Here Come the Fake Videos, Too.” *The New York Times*. March 4, 2018. nytimes.com.

[10] Adam Alter. *Irresistible: The Rise of Addictive Technology and the Business of Keeping Us Hooked*. Penguin Books. 2018.

ExternalURL
https://www.seattleu.edu/ethics-and-technology/viewpoints/deepfakes-and-the-val...

## Notes

Authors: Nathan Colaner and Michael J. Quinn, February 10, 2020.

## Rights

## Resource Type

Essay

## Topics

Public Well-being
Social Responsibility
Bias in Research
Data Management
Research Misconduct
Falsification

## Discipline(s)

Computer, Math, and Physical Sciences
Computer Sciences