Regulatory
Studies Center
THE GEORGE WASHINGTON UNIVERSITY

Improving regulatory policy through research, education, & outreach

Possibilities & Perils of Deepfake Technology

By: Dylan Desjardins | April 6, 2022

While the potential for deepfake technology to disrupt society have been clear for years, its growing accessibility and high-profile instances of use have increased concerns over their capabilities. In the first of a planned series discussing deepfake regulation, this post summarizes some of the most significant costs and benefits of the technology.

In mid-March, a video appeared online of Ukrainian President Zelensky apparently declaring his surrender to Russia. Social media quickly declared it false and took it down, but it serves to illustrate concerns over deepfake technology, which have steadily increased in recent years. The Government Accountability Office defines a deepfake as "a video, photo, or audio recording that seems real but has been manipulated with AI," typically through synthetization or replacement of speech and faces. As deepfake technology becomes more accessible, the total number of deepfakes released online has steadily accelerated. Once accessible only to a select few, computer applications for deepfake creation paired with tutorials are now broadly available online, allowing anyone with basic computer skills to attempt to create a deepfake. There are limits to these deepfakes' persuasiveness—for example most applications still require many "training images" to create a realistic product, making convincing deepfakes of those without a public persona difficult. But the remaining barriers to widescale convincing deepfakes may also decrease in the future, so the time is right to consider what to do about them.

Potential Harms

While the Zelensky deepfake was immediately derided as clunky and unrealistic, regulators need to examine more than a global news story to assess the total impact of deepfakes.

Government institutions of all levels stand to be harmed by deepfakes. There is potential that false information imparted through a fake video (or confusion over whether a video is real) could significantly disrupt normal operations. A faked video of President Zelensky giving orders could lead to actions directly counter to the nation's interest—for example causing confusion over whether military ought to lay down arms or leave the country. Other institutions (major businesses, advocacy groups) could face similar attacks.

A second group conceivably harmed is the audience. Viewers could suffer harm if they make damaging decisions or face harmful outcomes based on false information. For example, if a false video persuaded

viewers to vote in a way they normally wouldn't or against their own interests, that would be harmful to them. This also suggests that if a deepfake video gains traction, the public at large (even non-viewers) could suffer harms. If a deepfake video incited viewers to violence against a segment of the public that did not view the video, the whole society would incur costs.

Additionally, the prevalence of deepfakes, rather than any one deepfake, could have psychological effects on populations, making it harder for them to discern true information and less able to access accurate information to inform their own decisions and well-being. A recent paper notes that statements warning about the existence of deepfakes induced participants of a study to believe that even real videos they were shown were likely fake. This suggests that the danger is not just a deepfake video spreading in society, but society believing there are no meaningful checks on deepfakes within their information ecosystem.

Beyond these institutional and societal harms, there are also individual harms to consider. Individuals in deepfakes are harmed when the existence of the video ascribes false speech or actions to them. Viewers of the deepfake may regard these individuals unfavorably after being exposed to their likeness in the deepfake, perhaps because the faked video shows them engaging in a disapproved activity. This can lead to reputational or legal harms, or in some instances may even facilitate physical violence. Arguably, the vast majority of deepfakes today levy the greatest harms on these individuals: non-consensual deepfake pornography makes up the vast majority of deepfakes online, with Deeptrace estimating they made up 96% of deepfake videos in 2019.

One deepfake will not necessarily trigger only one of the harms above; it could result in several. While a deepfake that manipulates a celebrity image might just damage that celebrity, a deepfake that shows a president declaring war (or in the case of Ukraine surrendering) likely affects the individual president's reputation, the public's trust, and international relations.

When we think of this risk, we could think of a few well-placed deepfakes wreaking havoc on society: perhaps a foreign power throwing an election into disarray through a well-placed deepfake of a candidate. But when the technologies to produce and distribute deepfakes are widely accessible, more individual harms could result. As just one example, imagine the effect deepfakes could have in everyday divorce courts and custody proceedings when one side can produce a realistic-looking video of their former partner verbally abusing their child.

Potential Benefits

While they rarely elicit headlines or inspire movie scripts, deepfakes have also been developed for socially beneficial purposes. Deepfakes can help make translation and dubbing of videos easier and more efficient; a movie could be translated into myriad languages in a way that makes the actor appear to actually speak the given dialect. Deepfakes are also being used for various entertainment and commercial purposes, such as having shoppers try clothes they are considering on life-like avatars. Museums might consider using deepfake technology to create convincing videos of historical figures speaking or moving, allowing for interactive and engaging exhibits.

Research and development of deepfakes might also lead to unrelated breakthroughs in related technologies. Scientists purportedly studying one type of technology have a history of advancements in initially unforeseen avenues, a classic example being the invention of the microwave by an engineer studying radar-related technology. Related to this, there may be currently untapped benefits to deepfake technology.

Future posts will examine what actions federal agencies are taking now to combat deepfake harms, and how their mandates or activities might be expanded in the future. If you have any thoughts on these topics, feel free to share them with ddesjar2@gwu.edu.