Congressional Research Service
Informing the legislative debate since 1914

IN FOCUS

October 14, 2019

Deep Fakes and National Security

"Deep fakes"—a term that first emerged in 2017 to describe realistic photo, audio, video, and other forgeries generated with artificial intelligence (AI) technologies—could present a variety of national security challenges in the years to come. As these technologies continue to mature, they could hold significant implications for congressional oversight, U.S. defense authorizations and appropriations, and the regulation of social media platforms.

How Are Deep Fakes Created?

Though definitions vary, deep fakes are most commonly described as forgeries created using techniques in machine learning (ML)—a subfield of AI—especially generative adversarial networks (GANs). In the GAN process, two ML systems called neural networks are trained in competition with each other. The first network, or the generator, is tasked with creating counterfeit data—such as photos, audio recordings, or video footage—that replicate the properties of the original data set. The second network, or the discriminator, is tasked with identifying the counterfeit data. Based on the results of each iteration, the generator network adjusts to create increasingly realistic data. The networks continue to compete—often for thousands or millions of iterations—until the generator improves its performance such that the discriminator can no longer distinguish between real and counterfeit data.

Though media manipulation is not a new phenomenon, the use of AI to generate deep fakes is causing concern because the results are increasingly realistic, rapidly created, and cheaply made with freely available software and the ability to rent processing power through cloud computing. Thus, even unskilled operators could download the requisite software tools and, using publically available data, create increasingly convincing counterfeit content.

How Could Deep Fakes Be Used?

Deep fake technology has been popularized for entertainment purposes—for example, social media users inserting the actor Nicholas Cage into movies in which he did not originally appear and a museum generating an interactive exhibit with artist Salvador Dalí. Deep fake technologies have also been used for beneficial purposes. For example, medical researchers have reported using GANs to synthesize fake medical images to train disease detection algorithms for rare diseases and to minimize patient privacy concerns.

Deep fakes could also be used for nefarious purposes. State adversaries or politically motivated individuals could release falsified videos of elected officials or other public figures making incendiary comments or behaving inappropriately. Doing so could, in turn, erode public trust, negatively affect public discourse, or even sway an election. Indeed, the U.S. intelligence community concluded that Russia engaged in extensive influence operations during the 2016 presidential election to "undermine public faith in the U.S. democratic process, denigrate Secretary Clinton, and harm her electability and potential presidency." In the future, convincing audio or video forgeries could potentially strengthen similar efforts.

Deep fakes could also be used to embarrass or blackmail elected officials or individuals with access to classified information. Already there is evidence that foreign intelligence operatives have used deep fake photos to create fake social media accounts from which they have attempted to recruit Western sources. Some analysts have suggested that deep fakes could similarly be used to generate inflammatory content—such as convincing video of U.S. military personnel engaged in war crimes—intended to radicalize populations, recruit terrorists, or incite violence.

In addition, deep fakes could produce an effect that professors Danielle Keats Citron and Robert Chesney have termed the "Liar's Dividend"; it involves the notion that individuals could successfully deny the authenticity of genuine content—particularly if it depicts inappropriate or criminal behavior—by claiming that the content is a deep fake. Citron and Chesney suggest that the Liar's Dividend could become more powerful as deep fake technology proliferates and public knowledge of the technology grows.

Some reports indicate that such tactics have already been used for political purposes. For example, political opponents of Gabon President Ali Bongo asserted that a video intended to demonstrate his good health and mental competency was a deep fake, later citing it as part of the justification for an attempted coup. Outside experts were unable to determine the video's authenticity, but one expert noted, "in some ways it doesn't matter if [the video is] a fake… It can be used to just undermine credibility and cast doubt."

How Can Deep Fakes Be Detected?

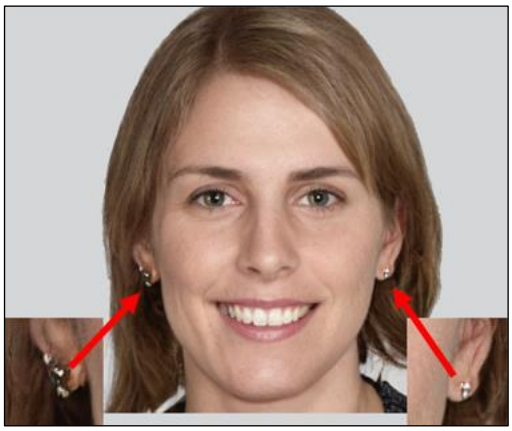
Today, deep fakes can often be detected without specialized detection tools. However, the sophistication of the technology is rapidly progressing to a point at which unaided human detection will be very difficult or impossible. While commercial industry has been investing in automated deep fake detection tools, this section describes the U.S. government investments at the Defense Advanced Research Projects Agency (DARPA).

DARPA currently has two programs devoted to the detection of deep fakes: Media Forensics (MediFor) and Semantic Forensics (SemaFor). MediFor is developing algorithms to automatically assess the integrity of photos

and videos and to provide analysts with information about how counterfeit content was generated. The program has reportedly explored techniques for identifying the audio-visual inconsistencies present in deep fakes, including inconsistencies in pixels (digital integrity), inconsistencies with the laws of physics (physical integrity), and inconsistencies with other information sources (semantic integrity). MediFor received $17.5 million in FY2019 and is slated to receive $5.3 million in FY2020 as the program begins to transition to operational commands and the intelligence community.

Similarly, SemaFor seeks to develop algorithms that will automatically detect, attribute, and characterize (i.e., identify as either benign or malicious) various types of deep fakes. This program will catalog semantic inconsistencies—such as the mismatched earrings seen in the GAN-generated image in **Figure 1**, or unusual facial features or backgrounds—and prioritize suspected deep fakes for human review. Both SemaFor and MediFor are intended to improve defenses against adversary information operations.

Figure 1. Example of Semantic Inconsistency in a GAN-Generated Image



Source: https://www.darpa.mil/news-events/2019-09-03a.

Policy Considerations

Some analysts have noted that algorithm-based detection tools could lead to a cat-and-mouse game, in which the deep fake generators are rapidly updated to address flaws identified by detection tools. For this reason, they argue that social media platforms—in addition to deploying deep fake detection tools—may need to provide a means of labeling and/or authenticating content. This could include a requirement that users identify the time and location at which the content originated or that they label edited content as such.

Other analysts have expressed concern that regulation of deep fake technology could impose undue burden on social media platforms or lead to unconstitutional restrictions on free speech and artistic expression. These analysts have suggested that existing law is sufficient for managing the malicious use of deep fakes. Some experts have asserted that responding with technical tools alone will be insufficient and that instead the focus should be on the need to educate the public about deep fakes and minimize incentives for creators of malicious deep fakes.

Potential Questions for Congress

- Does the Department of Defense, the Department of State, and the intelligence community have adequate information about the state of foreign deep fake technology and the ways in which this technology may be used to harm U.S. national security?
- How mature are DARPA's efforts to develop automated deep fake detection tools? What are the limitations of DARPA's approach, and are any additional efforts required to ensure that malicious deep fakes do not harm U.S. national security?
- Are federal investments and coordination efforts, across defense and nondefense agencies and with the private sector, adequate to address research and development needs and national security concerns regarding deep fake technologies?
- How should national security considerations with regard to deep fakes be balanced with free speech protections, artistic expression, and beneficial uses of the underlying technologies?
- Should social media platforms be required to authenticate or label content? Should users be required to submit information about the provenance of content? What secondary effects could this have for social media platforms and the safety, security, and privacy of users?
- To what extent and in what manner, if at all, should social media platforms and users be held accountable for the dissemination and impacts of malicious deep fake content?
- What efforts, if any, should the U.S. government undertake to ensure that the public is educated about deep fakes?

CRS Products

CRS Report R45178, *Artificial Intelligence and National Security*, by Kelley M. Sayler

CRS In Focus IF10608, *Overview of Artificial Intelligence*, by Laurie A. Harris

CRS Report R45142, *Information Warfare: Issues for Congress*, by Catherine A. Theohary

Other Resources

Office of the Director of National Intelligence, *Background to "Assessing Russian Activities and Intentions in Recent US Elections"*, January 6, 2017, https://www.dni.gov/files/documents/ICA_2017_01.pdf

Kelley M. Sayler, Analyst in Advanced Technology and Global Security
Laurie A. Harris, Analyst in Science and Technology Policy

Disclaimer

This document was prepared by the Congressional Research Service (CRS). CRS serves as nonpartisan shared staff to congressional committees and Members of Congress. It operates solely at the behest of and under the direction of Congress. Information in a CRS Report should not be relied upon for purposes other than public understanding of information that has been provided by CRS to Members of Congress in connection with CRS's institutional role. CRS Reports, as a work of the United States Government, are not subject to copyright protection in the United States. Any CRS Report may be reproduced and distributed in its entirety without permission from CRS. However, as a CRS Report may include copyrighted images or material from a third party, you may need to obtain the permission of the copyright holder if you wish to copy or otherwise use copyrighted material.