Defending Against Deep Fakes

Through Technological Detection, Media Literacy, and Laws and Regulations

Alina Alimova

Alina Alimova is currently pursuing her MA in Security Policy Studies at the Elliott School. Her concentration is on Transnational Security and she has focused on cybersecurity issues throughout her time at Elliott. Her favorite topics include information operations, disinformation campaigns, and critical infrastructure security. She has a Bachelor's degree in International Studies from American University and she has previously interned at the Atlantic Council Eurasia Center and the Alliance for Securing Democracy at the German Marshall Fund. Alina is currently interning at a due diligence firm using her Russian and English skills in her research.

ABSTRACT

Deep fakes, synthesized media created by computer networks, have become a prominent part of the national security conversation over the past few years. They are a prime example of a dual-use technology that can be used for both civil and military purposes. Deep fakes can be used to spread disinformation and propaganda, confuse the population, commit crimes, and even undermine democracy. However, deep fake technology can also be used for non-malicious purposes such as film making, therapy, and education. This paper explores deep fakes and how the US can defend against their malicious manifestations. The best way to defend against deep fakes is through a holistic approach based on technological detection, media literacy to protect against disinformation, and comprehensive laws and regulations. Because deep fakes affect all levels of society, partnerships and multi stakeholder engagement will be at the center of effective policy. The US defense community will have to adopt a novel approach that goes outside the traditional confines of security. This paper presents three interconnected recommendations: creating partnerships with technology companies capable of developing software that can track deep fakes; increasing focus on media literacy programs; and addressing legal and regulatory challenges posed by deep fakes. The goal should be to create a population that is resilient to both foreign and domestic deep fakes. Overall, this is a complex and multifaceted issue that requires collaboration from a variety of stakeholders.

BACKGROUND

A deep fake is an artificial intelligence (AI)-based technology used to produce synthetic content. Creating a deep fake requires two neural networks, which are computational learning systems that are programmed to translate data from one form into another.[1] The two neural networks work together in a feedback loop to create a deep fake. First, one neural network generates fake content based on a database of real content. Then, the second neural network compares the synthetic content to existing real content.[2] This ultimately creates a feedback loop that continuously improves synthetic content, including text, audio, video and images.

Deep fakes are increasingly important within the national security sphere in recent years. While AI is not a new phenomenon, Tim Hwang asserts that deep fakes are becoming a 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."[3] With the growing dissemination of deep fakes, fake content will become increasingly prominent in cyberspace. Additionally, Hwang argues that deep fakes created by more sophisticated actors are going to become a greater threat over time.[4] While contemporary deep fakes can be identified fairly easily, the increasingly realistic content is likely to pose a greater threat to national security, including to elections, markets, and the military. Policymakers believe that large-scale information operations will integrate machine learning and AI, such as those launched during the 2016 U.S. presidential election.[5] This poses a threat to national security since it affects the population, erodes democracy, and decreases trust in government on a larger scale than traditional disinformation campaigns.

In terms of impact on society, deep fakes can undermine democracy, exacerbate fissures in societies, and erode trust in governments and institutions. Long-term effects could result in truth decay, long-term apathy, and a general erosion of truth or any interest in identifying it.[6] Conversely, individuals could successfully deny the authenticity of genuine content and claim that real footage is in fact a deep fake, in what experts call the "liar's dividend."[7] This denial can lead to individuals being exonerated of crimes they did commit, and the liar's dividend could become more powerful as technology proliferates and public awareness of deep fakes grows.[8]

There have already been instances of adversarial governments using deep fakes for political gain. In May 2018, Belgian political party Socialistische Partij Anders created a deep fake of U.S. President Trump commenting on his withdrawal from the Paris Climate agreement and urging other countries to do so as well.[9] In May 2019 a confession of a Malaysian political aid admitting to an affair with the Economics Affairs Minister was widely thought to be a

deep fake.[10] The EU-funded East StratCom Task Force has also taken note of Russian trolls experimenting with the use of deep fakes for disinformation.[11] These incidents demonstrate that even current deep fake technology is exploited for political or personal gain.

Due to the increasing risks to both civil society and national security, the U.S. defense community must adopt a set of holistic policies to defend against deep fakes. The U.S. Department of Defense's (DOD) Joint Artificial Intelligence Center (JAIC) would be responsible for research and policy implementation since deep fakes fall under their jurisdiction.[12] Additionally, the defense community has sufficient resources to trace a deep fake both through technology and accessibility to other resources, namely communication with other sectors and through public-private partnerships. The remainder of this paper explores three approaches to defending against deep fakes, specifically technological detection, media literacy programs to counter disinformation, and laws and regulations. Finally, the paper outlines recommendations for policymakers and the JAIC, primarily focused on cultivating partnerships to defend against deep fakes.

TECHNOLOGICAL DETECTION

Developing technological capabilities to quickly identify deep fakes is essential for a successful defense because it will allow governments to respond in a timely manner. The DOD is currently developing two programs devoted to the detection of deep fakes: Media Forensics (MediFor) and Sematic Forensics (SemaFor). The development of both programs is under the supervision of the JAIC.

MediFor develops algorithms to "automatically assess the integrity of photos and videos and to provide analysts with information about how counterfeit content was generated."[13] The program relies on a three-tiered framework of information including digital integrity, physical integrity, and semantic integrity.[14] MediFor is expected to be operational and in use by the intelligence community by 2021. SemaFor is developing algorithms that will automatically detect deep fakes based on semantic inconsistencies—such as mismatched earrings or unusual facial features.[15] SemaFor relies on the semantic level of the internet, meaning relying on the human use of the internet. Both MediFor and SemaFor are intended to improve defenses against adversarial information operations.

AI scientists developed another technology that focuses on using a convolutional neural network (CNN) in conjunction with a recurrent neural network (RNN), which allows the program to determine whether a video has been manipulated or not. A CNN is a deep learning algorithm that analyzes images.[16] An RNN is a type of neural network which uses sequential data

or time series data and is commonly used in applications such as speech recognition, and natural language translation and processing.[17] In a 2018 study, Guera and Delp demonstrated that a CNN and an RNN feedback loop is an effective way to repeatedly recognize deep fakes.[18] Essentially, this method relies on teaching AI to recognize deep fakes the way a human would, through small inconsistencies and errors in visual and linguistic content. This method is another promising avenue for technological detection of deep fakes.

Private companies also have in place several content monitoring systems. For example, YouTube has a program called SystemID, which identifies copyright issues from a database of content submitted by users; Adobe has the Content Authenticity Initiative which creates content attribution; and Microsoft has the AETHER Media Provenance (AMP) system that ensures the authentication of media.[19] These initiatives demonstrate a willingness by private companies to ensure the quality and authenticity of content on their platforms. This, in turn, provides a good base for future collaboration with government programs and public-private partnerships.

However, there are several challenges with technological detection. First, deep fake technology is continuously developing, making it difficult for detection technology to keep up. Additionally, attribution in cyberspace is often difficult and time consuming; therefore, tracing a deep fake to a particular actor may be difficult, even with the DoD programs under development. Furthermore, the current media landscape favors the manipulators, since technology is becoming more readily available and the public calls into question even official sources' claims. In other words, even with current disinformation campaigns, the claims by official news sources and experts are not trusted and the population continues to believe the false information that malicious actors spread. This issue is only likely to become more prominent as the technology and utilization of deep fakes commodifies. For these reasons, it is essential for the U.S. government to take action to counter deep fakes.

MEDIA LITERACY PROGRAMS

Deep fakes are closely related to traditional disinformation campaigns, therefore protection against them should be rooted in defending against disinformation. The best way to protect against disinformation is to create a resilient population through education and media literacy programs. According to a 2018 Disinformation Resilience Index (DRI) report by Prism UA, countries that scored highest on resilience, namely Estonia and Lithuania, were also the most resilient to Russian disinformation.[20] While the DRI only focused on Eastern and Central European countries, they were still able to demonstrate that systemic responses and education about disinformation created a more resilient civil society. Media literacy programs should focus on creating a society that is

aware of the threat of deep fakes and able to recognize them.

The United States and many Western European countries, however, were woefully unprepared to defend against disinformation campaigns in 2016 and are still unprepared today. In 2017, the U.S. intelligence community released a report that showed significant state-sponsored disinformation during the 2016 U.S. presidential elections.[21] Further, in 2020, the COVID-19 crisis demonstrated that the U.S. public is still highly susceptible to disinformation, both domestic and international. For example, many people doubted the effectiveness of masks, social distancing, and even the reality of the pandemic, leading to an increased spread of the virus.[22] The U.S. population has been overwhelmed with news containing fake reports and misinformation, which is difficult to digest. This inundation led to confusion, eroding trust in government and resulting in more deaths.[23] Therefore, countering the spread of disinformation is essential for national security.

With the proliferation of deep fakes, all of these issues are likely to be exacerbated. Matt Chessen argues that MADCOMs—the integration of machine learning into computational propaganda—will lead to highly personalized propaganda based on an individual's internet history, browsing habits, and behavior.[24] This will give propagandists radically enhanced capabilities to manipulate human minds, resulting in even more chaos than in current disinformation campaigns. While deep fakes can currently be recognized fairly easily,[25] as technology improves and learns from itself, the need for a resilient population is only going to grow.

Media literacy programs should focus on educating citizens to identify deep fakes as well as separating them from the truth, and there are several programs already in place. For example, MediaWise is a partnership between Google, Poynter, Stanford University, and the Local Media Association which teaches youth to differentiate fact from fiction. Facebook and Reuters have partnered to publish a media literacy course for manipulated media.[26] The Washington Post, CNN, and University of Cambridge have each published their own guides to recognizing deep fakes.[27] These are promising steps forward for the future development of media literacy programs.

LEGISLATION AND REGULATION

Legislation concerning deep fakes focuses on two main routes: civil and criminal litigation. Both, however, are notoriously difficult given the novel nature of deep fakes and inherent challenges of video evidence in courts. The following discussion provides a broad overview of the current civil and criminal litigation featuring deep fakes, the challenges associated with them, and efforts to improve legislation surrounding deep fakes.

First, civil litigation primarily focuses on prosecuting the creation of deep

fakes as the crime. When prosecuting the creation of deep fakes itself, lawyers have argued cases based on defamation and copyright laws.[28] Additionally, in June 2019, Virginia became the first state to officially ban deep fakes; Texas followed in September 2019, and California in October 2019. However, since deep fakes can be used for non-malicious purposes, legislation must be crafted in a way that does not infringe on fundamental freedoms. For example, in Italy, the creators of a deep fake of the Prime Minister making vulgar comments and gestures claimed that their video was satire when challenged.[29] This demonstrates that civil courts will likely have to consider the creation of deep fakes on a case-by-case basis.

Criminal litigation focuses on prosecuting the individuals using deep fakes to commit another crime, such as fraud. In other words, the deep fake in criminal cases is not the crime, but rather the tool used to commit a crime.[30] However, because evidentiary standards in criminal cases are stricter than in civil cases, there are still many challenges with using deep fakes in criminal proceedings. The prosecution in criminal cases involving deep fakes will have to prove the authenticity of the deep fake and the fact that it was created by the defendant.[31]

However, current commercially available technology has difficulty proving the origin and authenticity of audiovisual content. In 2018, Rössler et al. tested over half a million edited images with machine learning algorithm XceptionNet and found that it was effective at identifying manipulated images, but that the same algorithm could be used to create manipulated images as well.[32] Additionally, it is even more difficult to prove the authenticity of lower-quality videos, whether they are deep fake or not. This means that eyewitness testimony and other corroborating information is needed for a solid evidentiary base in future legal proceedings.[33]

On the other hand, there is the issue of the "liar's dividend," where an individual can claim that a real video is a deep fake or manipulation. Citron and Chessney claim that as education and awareness of deep fakes improve, the public may be more inclined to believe a lie about a genuine video being a fake. In other words, "a skeptical public will be primed to doubt the authenticity of real audio and video evidence."[34] This will be a challenge for U.S. legislators, because it can lead to higher burdens of proof and may even result in a criminal being found not guilty because of the "liar's dividend." All of these aspects of criminal proceedings show that the U.S. courts will need to adapt to prosecuting crimes that are committed in the future, as deep fake technology becomes more commercially available.

Regulation of information in the cyber domain is notoriously difficult and faces a multitude of challenges. Still, effectively curbing the spread of deep fakes on social media is essential in defending against them. Experts fear that algorithm-based detection tools could lead to a "cat and mouse" game between

two rapidly evolving technologies.[35] Therefore, it will be essential for social media companies to step in and regulate manipulated and synthetic content on their platforms. Some steps have already been taken towards that goal. In January 2020, Facebook updated its policy on synthetic media, committing to removing synthetic content from its website.[36] In February 2020, Twitter followed suit and shared its guidelines on synthetic media, which primarily focus on labeling and removing harmful fake content.[37] However, these policies exclude parody, satire, or content manipulated for cosmetic reasons.

Congress has also taken action to regulate deep fakes. In October 2019, the House and Senate passed the Deep Fake Report Act of 2019. This act aims to regulate deep fakes at the federal level and "requires the Science and Technology Directorate in the Department of Homeland Security to report at specified intervals on the state of digital content forgery technology."[38] While this is a very general and sweeping piece of legislation, it establishes a base for future regulatory legislation.

However, legislating and regulating deep fakes presents its own challenges. Primarily, policymakers will have to grapple with the issue of free speech. Since deep fakes can be used for art, filmmaking, therapy, and education, a blanket ban on all deep fakes will likely face criticism from congress and civil liberties organizations.[39] Additionally, courts will likely have to adjust evidentiary standards for media and introduce technological detection tools in courtrooms. This is likely to be costly, require a multitude of resources and will be challenging for local and regional courts. Social media companies will have to become more vigilant to synthetic content on their platforms and ensure higher regulatory standards for harmful deep fakes on their systems.

POLICY IMPLICATIONS, RECOMMENDATIONS, AND CONCLUSION

The policy implications of deep fakes are multifold. First, without an effective defense policy against deep fakes, the U.S. risks undermining its democracy and succumbing to disinformation to an even greater extent. Deep fakes can potentially lead to unintended escalation; for example, a deep fake of a politician could provoke an adversary and cause a political conflict. In a worst-case scenario, politicians and military decision-makers could take action based on false information spread by deep fakes. Deep fakes and computational propaganda risk undermining U.S. democracy, exploiting divisions in society, and undermining public trust in government. It is essential for U.S. policymakers to re-establish trust domestically. Otherwise, the country will face significant challenges, as demonstrated by the COVID-19 crisis. Additionally, regulatory and legal systems have not fully caught up to the technology, making prosecuting crimes more difficult. These issues pose a significant risk to U.S. national security.

As outlined above, these are complex issues that will require cooperation from multiple stakeholders for effective redressal. The solution to these issues requires a set of holistic policies spearheaded by the national defense community, but implemented by partnerships with other stakeholders. Since this national security issue is closely tied to the public sector, public-private partnerships will play a crucial role in enhancing the defense against deep fakes. Since the JAIC is the entity responsible for overseeing the development of technologies to detect deep fakes, this agency should focus on countering deep fakes through technological detection. Additionally, the JAIC should establish relationships with critical stakeholders in the private sector, academia, non-governmental organizations, and grassroots movements to create effective media literacy programs against deep fakes. Finally, the JAIC should facilitate the formation of partnerships between representatives of legislative and regulatory communities to create effective laws and regulations to prosecute civil and criminal issues related to deep fakes.

First, the JAIC should create partnerships with technology companies capable of developing software that can track deep fakes. However, one of the biggest challenges with technological detection is that deep fake creation tools are rapidly developing, which means that detection software must evolve continuously as well. One recommendation to address challenges posed by the rapid technological development of deep fakes is to commodify detection, as Hwang argued.[40] The JAIC should make detection technologies more readily available for commercial use, such as in social media companies, news and media organizations, and companies that control critical infrastructure. This could include giving private companies subsidies to make their software public or cooperating with private firms to create publicly available software. Taking these steps would ease detection and allow for faster action against deep fakes.

Second, the JAIC should increase focus on fostering media literacy programs. Since technology alone may not be enough to detect and remove deep fakes, an educated society is the best defense against deep fakes' potential for disinformation threats. If the society is resilient to international and domestic propaganda and disinformation posed by deep fakes, their effectiveness will significantly decrease. For this reason, partnerships that focus on deep fake education detection are crucial. The JAIC should rely on programs previously established by academia and news organizations. The JAIC should establish partnerships with grassroots organizations as well as trusted local newspapers and journalists to enhance their media literacy programs' effectiveness. Ultimately, the goal should be to create a population that is resilient to both foreign and domestic deep fakes.

Finally, U.S. policymakers must address the legal and regulatory challenges deep fakes present. The U.S. government will need to work closely with the legal system to establish a fair and balanced legal process to prosecute crimes

related to deep fakes. However, the JAIC and any other government agencies should only act in an advisory and expert capacity to ensure that there is no government influence on the legal process. Furthermore, the U.S. legal system will have to grapple not only with the novelty of deep fake technology but also the difficulties of legislating a technology that could potentially face opposition by civil rights and first amendment protection groups. The JAIC should also offer partnerships to provide courts with cheaper and more accessible technological detection capabilities. These steps are just the beginning to effectively legislating and prosecuting deep fakes.

In terms of regulatory partnerships, the JAIC can work with social media companies to enhance their detection and regulatory capabilities. Facebook, which demonstrated a willingness to take steps towards removing deep fakes from its platform, is likely to be open to partnerships that enhance its capabilities. Other major social media companies should use similar approaches. Congress should build on the Deepfake Report Act of 2019 to create a federal basis for regulating and detecting deep fakes.

Importantly, none of these recommendations exist in a vacuum. In order to effectively protect against deep fakes, policymakers must undertake a holistic approach. Such an approach will involve a multitude of partners and stakeholders, and is likely to require significant initial resources from the government. If the United States can effectively protect against the threat posed by deep fakes, the initial investment in partnerships will pay for itself in the future. Overall, this is a complex and multifaceted issue that requires collaboration from a variety of stakeholders.

ENDNOTES

1 "Neural Network,"DeepAI, May 17, 2019, https://deepai.org/machine-learning-glossary-and-terms/neural-network.

2 D. J. Pangburn, "You've Been Warned: Full Body Deepfakes Are the next Step in AI-Based Human Mimicry," Fast Company, September 21, 2019, https://www.fastcompany.com/90407145/youve-been-warned-full-body-deepfakes-are-the-next-step-in-ai-based-human-mimicry.

3 "Deep Fakes and National Security" Congressional Research Service (August 26, 2020), https://crsreports.congress.gov/product/pdf/IF/IF11333.

4 Tim Hwang, "Deepfakes: A Grounded Threat Assessment", Center for Security and Emerging Technology (July 2020):1-30, https://cset.georgetown.edu/research/deepfakes-a-grounded-threat-assessment/.

5 Hwang, "Deepfakes," 2.

6 Emilia Anna Porubcin, "Seeing Is No Longer Believing: Deepfakes, Cheapfakes and the Limits of Deception - CIAO," International Centre for Defense and Security (December 2019): 1-21, http://www.ciaonet.org.proxygw.wrlc.org/record/65093?search=1

7 "Deep Fakes and National Security," 1.

8 Ibid., pg. 1.

9 "Belgian Socialist Party Circulates 'Deep Fake' Donald Trump Video," POLITICO, May 21, 2018, https://www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreement-belgian-

socialist-party-circulates-deep-fake-trump-video/

10 "Is the Political Aide Viral Sex Video Confession Real or a Deepfake? Malay Mail," accessed December 8, 2020, https://www.malaymail.com/news/malaysia/2019/06/12/is-the-political-aide-viral-sex-video-confession-real-or-a-deepfake/1761422

11 Nick Harding, "Video Nasties: Russia's Faked Broadcasts a New Threat to West," The Sunday Telegraph, May 27, 2018, https://www.pressreader.com/uk/the-sunday-telegraph/20180527/281797104665830

12 "About the JAIC - JAIC," accessed December 13, 2020, https://www.ai.mil/about.html

13 "Deep Fakes and National Security," 2.

14 Connor Collins, "DARPA Tackles Deepfakes With AI," GovernmentCIO Media and Research, March 11, 2019, https://governmentciomedia.com/darpa-tackles-deepfakes-ai

15 "Deep Fakes and National Security," 2.

16 Sumit Saha, "A Comprehensive Guide to Convolutional Neural Networks — the ELI5 Way," Medium, December 17, 2018, https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

17 IBM Cloud Education, "What Are Recurrent Neural Networks?" IBM, 14 September 2020, accessed January 6, 2021, https://www.ibm.com/cloud/learn/recurrent-neural-networks

18 D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), (February 14, 2019): 1–6. https://doi.org/10.1109/AVSS.2018.8639163

19 Ashish Jaiman, "Technical Countermeasures to Deepfakes," Medium, August 27, 2020, https://towardsdatascience.com/technical-countermeasures-to-deepfakes-564429a642d3

20 Viktor Denisenko, "Disinformation Resilience Index," Ukrainian Prism Foreign Policy Council, July 31, 2018, http://prismua.org/en/9065-2/

21 "Assessing Russian Activities and Intentions in Recent US Elections," Office of the Director of National Intelligence, January 6, 2017, https://www.dni.gov/files/documents/ICA_2017_01.pdf?utm_source=fbia

22 Bagherpour Amir and Ali Nouri, "COVID Misinformation Is Killing People," Scientific American, October 11, 2020, https://www.scientificamerican.com/article/covid-misinformation-is-killing-people1/

23 Fabio Tagliabue, Luca Galassi, and Pierpaolo Mariani, "The 'Pandemic' of Disinformation in COVID-19," SN Comprehensive Clinical Medicine 2, (August 1, 2020): 1287-1289, https://doi.org/10.1007/s42399-020-00439-1

24 Matt Chessen, "The MADCOM Future," Atlantic Council , September 26, 2017, https://www.atlanticcouncil.org/in-depth-research-reports/report/the-madcom-future/

25 Porcubin,"Seeing Is No Longer Believing," 5.

26 Ashish Jaiman, "Media Literacy, an Effective Countermeasure for Deepfakes," Medium, September 8, 2020, https://medium.com/the-innovation/media-literacy-an-effective-countermeasure-for-deepfakes-c6844c290857

27 Ibid,1.

28 Agnes E. Venema and Zeno Geradts, "Digital Forensics, Deepfakes, and the Legal Process," The SciTech Lawyer 16, no. 4 (Summer 2020), https://www.essentialresearch.eu/2020/07/07/deepfakes/

29 Ibid, 3.

30 Ibid., 3

31 Marie-Helen Maras and Alex Alexandrou, "Determining Authenticity of Video Evidence in the Age of Artificial Intelligence and in the Wake of Deepfake Videos," The International Journal of Evidence and Proof 23, no. 3 (July 1, 2019): 255–62, https://doi.org/10.1177/1365712718807226

32 Andreas Rössler et al., "FaceForensics: A Large-Scale Video Dataset for Forgery Detection in Human Faces," arXiv:1803.09179v1 [cs.CV], March 24, 2018, http://arxiv.org/abs/1803.09179

33 Maras and Alexandrou, "Determining Authenticity."

34 Robert Chesney and Danielle Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," California Law Review 107, no. 6 (2019): 1753–1820.

35 "Deep Fakes and National Security," 2.

36 Monika Bickert,"Enforcing Against Manipulated Media," About Facebook , January 6, 2020, https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/

37 "Synthetic and Manipulated Media Policy," accessed February 26, 2021, https://help.twitter.com/en/rules-and-policies/manipulated-media

38 Rob Portman, "S.2065 - Deepfake Report Act of 2019," Congress.gov, October 29, 2019, https://www.congress.gov/bill/116th-congress/senate-bill/2065

39 Porubcin, "Seeing Is No Longer Believing," 8.

40 Hwang, "Deepfakes: A Grounded," 25.