HIGH SCHOOL EDITION
Journal of Student Research
Volume 12 Issue 3 (2023)

Detection and Provenance: A Solution to Deepfakes?

Rishabh Jain

Woodward Academy

ABSTRACT

AI-enabled disinformation has rapidly spread throughout our society. Day by day, deepfakes are being used by adversaries to spread fake news and harm the credibility of political institutions. Inaction only lets the cumulative effect of all the disinformation grow, and potentially collapse order and trust in society. Therefore, the United States should pursue a two-pronged strategy against deepfakes composed of enhancing detection initiatives and implementing content provenance across the internet. This would prevent people from believing in the disinformation spread across the internet as well as fend off adversaries weaponizing deepfakes and targeting the public.

Scope of the Problem

Deepfakes are an issue that urgently requires policy action. They are a form of fake media in which video and audio are synthesized and used to "train" an artificial intelligence network to produce images and videos of something that is not real. Given the rapid pace of AI developments, this technology could easily spread to a variety of malicious state, nonstate, and individual actors.[1] The potential for destruction is endless — deepfake videos of politicians would swing elections, polarization through social media would be exacerbated by AI-enabled disinformation, governments caught up in deepfake scandals would cause citizen backlash, and increasing mistrust of information would cause even true information to be perceived as incorrect, a term named the "liar's dividend".[2][3] In the worst-case scenario, a bad actor uses deepfakes to incite nuclear weapon usage by creating a video of a global leader announcing their intent to start a war.[4] This type of escalation is plausible — Pakistan's defense minister threatened Israel after a false article about Israel threatening Pakistan with nuclear weapons spread.[5] Furthermore, the technology is readily available now; easy-to-use online services and tutorials grant ordinary citizens the power to create their own deepfakes, which could plunge the world into chaos if used maliciously.[6] We as a society are already seeing this - fake images of Donald Trump's arrest have spread throughout the internet, and a faked video of Elizabeth Warren declaring Republicans

[1] Robert Chesney and Danielle K. Citron, "Disinformation on Steroids: The Threat of Deep Fakes," Council on Foreign Relations, last modified October 16, 2018, accessed February 10, 2023, https://www.cfr.org/report/deep-fake-disinformation-steroids.
[2] Todd C. Helmus, "Artificial Intelligence, Deepfakes, and Disinformation: A Primer," RAND, last modified July 2022, accessed February 10, 2023, https://www.rand.org/pubs/perspectives/PEA1043-1.html.
[3] Chesney and Citron, "Disinformation on Steroids," Council on Foreign Relations.
[4] Jack Langa, "Deepfakes, Real Consequences: Crafting Legislation to Combat Threats Posed by Deepfakes," *Boston University Law Review* 101, no. 2 (March 2021): 771, accessed February 17, 2023, https://www.bu.edu/bulawreview/files/2021/04/LANGA.pdf.
[5] Langa, "Deepfakes, Real," 771.
[6] Charlotte Stanton, "How Should Countries Tackle Deepfakes?," Carnegie Endowment for International Peace, last modified January 28, 2019, accessed February 10, 2023, https://carnegieendowment.org/2019/01/28/how-should-countries-tackle-deepfakes-pub-78221.

should not be allowed to vote blew up on Twitter.[7] But, even worse, there is a risk that revisionist powers such as Russia and China use AI in ways that could potentially disrupt societal stability and sow disorder across the globe. Thus, there is a need for the United States to quickly and effectively marshal allies by controlling AI and providing norms for democracies.[8]

Policy Recommendations

The best strategy for the United States to tackle deepfakes is a two-pronged approach involving enhancing detection measures and content provenance. Enhancing detection measures would involve working with technology companies to obtain data repositories for scanning and analysis as well as to implement deepfake detection broadly across the nation.[9] Content provenance is effectively a watermark on an original photo or video embedded through metadata, which would assure the public that media has not been tampered with at all; implementing this nationwide would require working with not only technology companies but other organizations dedicated to provenance such as the Content Authenticity Initiative and the Coalition for Content Provenance and Authority (C2PA).[10] However, the power of the federal government is still incredibly relevant to both of these efforts. Investing in federal research and development organizations such as the Defense Advanced Research Projects Agency (DARPA) is critical to developing provenance standards and detection measures. Specifically, increasing investments in DARPA's MediFor program is crucial in the fight against deepfakes.[11]

One of the biggest issues the approach would resolve is the issue of deepfake arms racing. Currently, adversaries and threats to national security utilize something called generative adversarial networks. These are machine learning networks that can rapidly create a deepfake, and Russia and China have been deploying these as part of their broader disinformation strategies.[12] Despite attempts to improve detection measures, the arms race is still being won by our adversaries.[13] Status quo efforts to contain deepfakes are failing because the adversaries simply get better at fixing and improving the machine learning networks. This is why a comprehensive detection strategy is important — DARPA cooperation with the tech sector has the potential to unlock new detection mechanisms and install them nationwide.[14]

For example, DARPA's MediFor program is developing techniques to use artificial intelligence itself to detect deepfakes. This kind of automated detection algorithm is exactly why the U.S. should proceed with a more advanced and fully funded detection strategy because if MediFor's techniques were applied at large, adversaries would

[7] Naomi Nix and Isaac Stanley-Becker, "Fake images of Trump arrest show 'giant step' for AI's disruptive power," editorial, Washington Post, last modified March 22, 2023, accessed April 7, 2023, https://www.washingtonpost.com/politics/2023/03/22/trump-arrest-deepfakes/.

[8] Martijn Rasser et al., "The American AI Century: A Blueprint for Action," Center for a New American Security, last modified December 17, 2019, accessed February 10, 2023, https://www.cnas.org/publications/reports/the-american-ai-century-a-blueprint-for-action.

[9] Chesney and Citron, "Disinformation on Steroids," Council on Foreign Relations.

[10] Helmus, "Artificial Intelligence," RAND.

[11] Connor Collins, "DARPA Tackles Deepfakes With AI," GovCIO Media & Research, last modified March 11, 2019, accessed February 10, 2023, https://governmentciomedia.com/darpa-tackles-deepfakes-ai.

[12] Matthew Fecteau, "The Deep Fakes Are Coming," War Room, last modified April 23, 2021, accessed April 6, 2023, https://warroom.armywarcollege.edu/articles/deep-fakes/.

[13] Helmus, "Artificial Intelligence," RAND

[14] Chesney and Citron, "Disinformation on Steroids," Council on Foreign Relations.

start falling behind in the arms race.[15] Overall, existing adversarial deepfake disinformation operations and the relatively slow pace of detection investment mandate that the U.S. substantially reorients its strategy towards deepfake detection.

However, detection should not be the only way the U.S. tackles deepfakes. Due to the nature of the internet and social media, thousands of low-quality deepfakes can easily spread across the web. The issue with that is detection algorithms oftentimes only work with high-quality media or when one knows the origin of the deepfake.[16] While the U.S. can certainly wait for detection algorithm breakthroughs to get around this barrier, our response to deepfakes should not solely hinge on detection solutions. Enter content provenance, a tool to effectively and efficiently trace the origin and history of a piece of media on the internet. The C2PA's technical definition of content provenance is a standard that "lets content creators and editors disclose who created the content; how, when, and where it was created; and when and how it was edited throughout its life. In turn, content consumers can view those disclosures and verify authenticity. "[17] It is effectively an open-sourced "version history" of pieces of media on the internet. This kind of broad standard for digital content will be able to control widespread, low-quality misinformation on the internet by making sure web users know if the digital media they consume has been tampered with.

Having said that, the combination of both detection and provenance is essential to the fight against deepfake disinformation. Detection is particularly useful in combating specific, large threats such as adversarial disinformation because it is the most effective way to debunk disinformation the U.S. knows the origin and patterns of. Provenance is crucial to reduce the spread of disinformation on the internet created by localized, smaller actors because of its ease of use and ability to be implemented on social media platforms. Targeted legislation to fund detection efforts, work with social media companies, and implement provenance standards across the board is the optimal path for deepfake regulation.

Policy Alternatives

While federal detection and provenance initiatives are certainly an effective way to combat deepfakes, there are two other ways to neutralize them: private-sector action and counter-disinformation diplomacy through the Department of State's Global Engagement Center (GEC).

Simply letting the private sector tackle deepfakes seems like a viable option. Currently, big tech companies such as Google and Facebook are working on ways to keep fake news off their websites. Both of the aforementioned companies have been creating fake videos and open-sourcing them for others to develop deepfake detectors off of them?.[18] In fact, Facebook claims to have developed a detector that can track the origin of deepfakes.[19] The private sector's track record of a willingness to fight deepfakes and history of developing counter-disinformation strategies sounds promising, but there are numerous drawbacks. The largest one is that action by specific companies causes

[15] Collins, "DARPA Tackles," GovCIO Media & Research.

[16] John Donegan, "Content provenance is our best chance in the fight against deepfakes," ManageEngine Insights, last modified February 17, 2022, accessed April 7, 2023, https://insights.manageengine.com/artificial-intelligence/content-provenance-is-our-best-chance-in-the-fight-against-deepfakes/.

[17] Donegan, "Content provenance," ManageEngine Insights.

[18] Cade Metz, "Internet Companies Prepare to Fight the 'Deepfake' Future," The New York Times, last modified November 24, 2019, accessed April 19, 2023, https://www.nytimes.com/2019/11/24/technology/tech-companies-deepfakes.html.

[19] Jaclyn Diaz, "Facebook Researchers Say They Can Detect Deepfakes And Where They Came From," NPR, last modified June 17, 2021, accessed April 19, 2023, https://www.npr.org/2021/06/17/1007472092/facebook-researchers-say-they-can-detect-deepfakes-and-where-they-came-from.

migration to different platforms.[20] The companies who enact deepfake regulation will just drive bad actors to start sowing discord on other platforms, circumventing any efforts to restrict them. Migration has happened before — for example, the alt-right movement migrated from Twitter to Parler after perceiving that their posts were being censored on the former. Moreover, private sector developments simply would not be enough to send a clear signal to adversaries. Only a coherent, clear federal response would show adversaries the U.S. understands how to combat deepfakes broadly, while a private sector approach would be fragmented and incomplete. If adversaries perceived weakness in deepfake defenses, they would simply flood the information environment to overwhelm and defeat the aforementioned defenses. Only counter-disinformation tactics applicable at all levels would be able to provide a bulwark against deepfakes, and only the federal government has the authority and capability to implement that.

The GEC is a perfect example of a potentially useful federal government approach to deepfakes. It has experience in defeating adversarial disinformation, which makes it a particularly attractive option for combatting deepfakes. Historically, it has worked to expose Chinese disinformation that produces pro-Russia narratives about Ukraine. It does this by attempting to conduct public diplomacy to dispel myths and propaganda spread by adversaries. It also has previously worked with news outlets and tech companies to deliver information on fake news being spread.[21] Furthermore, it has experience in defeating disinformation that comes from Russia itself. It has exposed Russian disinformation tactics in Africa intended to win the continent over to its side as the war in Ukraine rages on.[22] With all that into account, the GEC seems like an obvious choice to use to combat deepfakes. However, this strategy has numerous structural flaws. The Department of State's Inspector General published a report that highlighted several issues within the GEC, including the lack of a clear mandate to coordinate counter-disinformation tactics, turf wars with other agencies, lack of communication, poor organizational structure, and hesitancy to expose propaganda due to the risk it would disrupt diplomatic negotiations.[23] The laundry list of problems makes the GEC a very unappealing option for such a severe issue that spans multiple dimensions of society. It only has useful experience in defeating adversaries and has no proven record of being able to control deepfakes over the internet, like content provenance can. It would be better to stick to programs with established detection algorithms, such as DARPA's MediFor, to defend against deepfakes rather than shift responsibility to a program with little accountability or experience against this novel form of disinformation. In summary, while the GEC is certainly one of the few sectors of the federal government dedicated to disinformation, it is nothing more than that, and the immense task of controlling a deepfake pandemic should not be placed in its hands.

---

[20] Kathleen Mary Carley, "A Political Disinfodemic," in *COVID-19 Disinformation: A Multi-National, Whole of Society Perspective*, ed. Ritu Gill and Rebecca Goolsby (Springer, Cham, 2022), pg. 13, accessed April 19, 2023, https://doi.org/10.1007/978-3-030-94825-2_1.

[21] Bill Gertz, "State Department works to counter Ukraine disinformation from China," The Washington Times, last modified April 7, 2022, accessed April 18, 2023, https://www.washingtontimes.com/news/2022/apr/7/state-department-working-debunk-chinese-disinforma/.

[22] Julian Pecquet, "US looks to expose Russian propaganda in Africa," The Africa Report, last modified May 25, 2022, accessed April 18, 2023, https://www.theafricareport.com/207268/us-looks-to-expose-russian-propaganda-in-africa/.

[23] Bill Gertz, "State Department watchdog gives failing grade to new counter-disinformation center," The Washington Times, last modified September 19, 2022, accessed April 18, 2023, https://www.washingtontimes.com/news/2022/sep/19/state-department-watchdog-gives-failing-grade-new-/.

Conclusion

Deepfakes are a national security issue that, without immediate action, will inevitably escalate and have serious consequences for society. Only U.S. leadership through a two-pronged strategy of detection and provenance can check back adversarial deepfakes and prevent spillover to the internet. While private sector action and diplomacy through the Global Engagement Center are good alternatives, this two-pronged strategy is the only way to comprehensively defend against deepfakes and avoid disjointed regulatory failures. It is the only way to implement broad standards that successfully deter AI-enabled disinformation operations and ward off the increasing polarization and instability that comes along with them.

References

———. "State Department works to counter Ukraine disinformation from China." The Washington Times. Last modified April 7, 2022. Accessed April 18, 2023. https://www.washingtontimes.com/news/2022/apr/7/state-department-working-debunk-chinese-disinforma/.

Carley, Kathleen Mary. "A Political Disinfodemic." In COVID-19 Disinformation: A Multi-National, Whole of Society Perspective, edited by Ritu Gill and Rebecca Goolsby, 1-24. Springer, Cham, 2022. Accessed April 19, 2023. https://doi.org/10.1007/978-3-030-94825-2_1.

Chesney, Robert, and Danielle K. Citron. "Disinformation on Steroids: The Threat of Deep Fakes." Council on Foreign Relations. Last modified October 16, 2018. Accessed February 10, 2023. https://www.cfr.org/report/deep-fake-disinformation-steroids.

Collins, Connor. "DARPA Tackles Deepfakes With AI." GovCIO Media & Research. Last modified March 11, 2019. Accessed February 10, 2023. https://governmentciomedia.com/darpa-tackles-deepfakes-ai.

Diaz, Jaclyn. "Facebook Researchers Say They Can Detect Deepfakes And Where They Came From." NPR. Last modified June 17, 2021. Accessed April 19, 2023. https://www.npr.org/2021/06/17/1007472092/facebook-researchers-say-they-can-detect-deepfakes-and-where-they-came-from.

Donegan, John. "Content provenance is our best chance in the fight against deepfakes." ManageEngine Insights. Last modified February 17, 2022. Accessed April 7, 2023. https://insights.manageengine.com/artificial-intelligence/content-provenance-is-our-best-chance-in-the-fight-against-deepfakes/.

Fecteau, Matthew. "The Deep Fakes Are Coming." War Room. Last modified April 23, 2021. Accessed April 6, 2023. https://warroom.armywarcollege.edu/articles/deep-fakes/.

Gertz, Bill. "State Department watchdog gives failing grade to new counter-disinformation center." The Washington Times. Last modified September 19, 2022. Accessed April 18, 2023. https://www.washingtontimes.com/news/2022/sep/19/state-department-watchdog-gives-failing-grade-new-/.

Helmus, Todd C. "Artificial Intelligence, Deepfakes, and Disinformation: A Primer." RAND. Last modified July 2022. Accessed February 10, 2023. https://www.rand.org/pubs/perspectives/PEA1043-1.html.

Langa, Jack. "Deepfakes, Real Consequences: Crafting Legislation to Combat Threats Posed by Deepfakes." Boston University Law Review 101, no. 2 (March 2021): 761-801. Accessed February 17, 2023. https://www.bu.edu/bulawreview/files/2021/04/LANGA.pdf.

Metz, Cade. "Internet Companies Prepare to Fight the 'Deepfake' Future." The New York Times. Last modified November 24, 2019. Accessed April 19, 2023. https://www.nytimes.com/2019/11/24/technology/tech-companies-deepfakes.html.

Nix, Naomi, and Isaac Stanley-Becker. "Fake images of Trump arrest show 'giant step' for AI's disruptive power." Editorial. Washington Post. Last modified March 22, 2023. Accessed April 7, 2023. https://www.washingtonpost.com/politics/2023/03/22/trump-arrest-deepfakes/.

Pecquet, Julian. "US looks to expose Russian propaganda in Africa." The Africa Report. Last modified May 25, 2022. Accessed April 18, 2023. https://www.theafricareport.com/207268/us-looks-to-expose-russian-propaganda-in-africa/.

Rasser, Martijn, Megan Lamberth, Ainikki Riikonen, Chelsea Guo, Michael Horowitz, and Paul Scharre. "The American AI Century: A Blueprint for Action." Center for a New American Security. Last modified December 17, 2019. Accessed February 10, 2023. https://www.cnas.org/publications/reports/the-american-ai-century-a-blueprint-for-action.

Stanton, Charlotte. "How Should Countries Tackle Deepfakes?" Carnegie Endowment for International Peace. Last modified January 28, 2019. Accessed February 10, 2023. https://carnegieendowment.org/2019/01/28/how-should-countries-tackle-deepfakes-pub-78221.