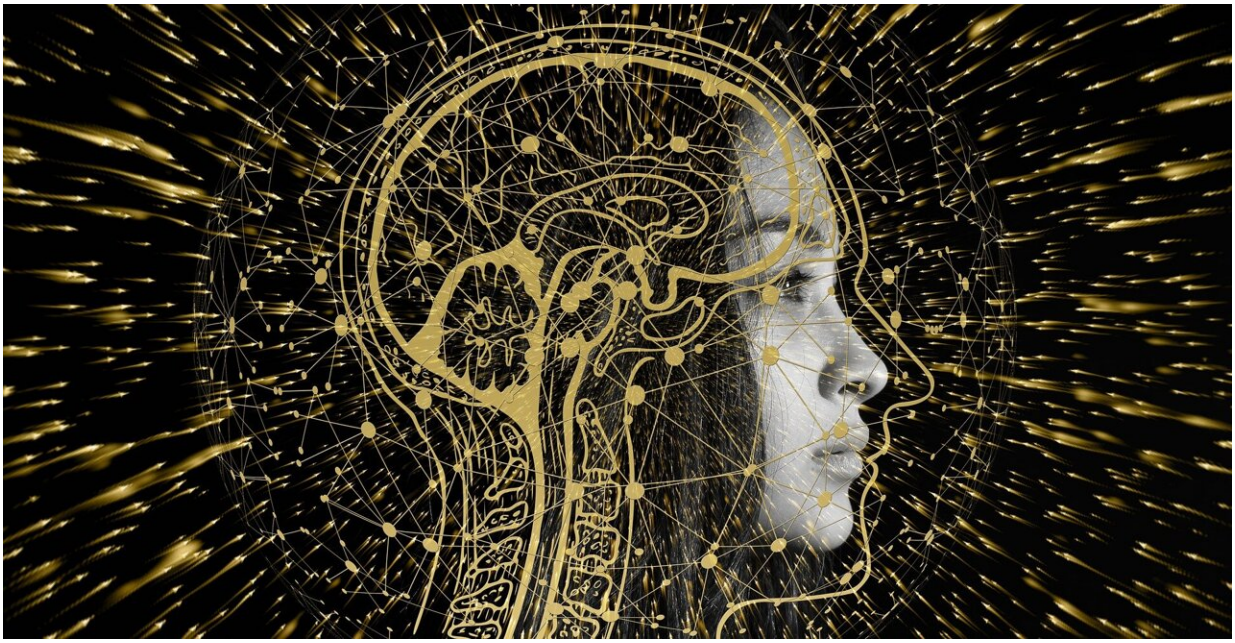
New report outlines recommendations for defending against deepfakes

January 17 2023, by Amanda Morris



Credit: Pixabay/CC0 Public Domain

Although most public attention surrounding deepfakes has focused on large propaganda campaigns, the problematic new technology is much more insidious, according to a new report by artificial intelligence (AI) and foreign policy experts at Northwestern University and the Brookings Institution.

In the new report, the authors discuss deepfake videos, images and audio

as well as their related security challenges. The researchers predict the technology is on the brink of being used much more widely, including in targeted military and intelligence operations.

Ultimately, the experts make recommendations to security officials and policymakers for how to handle the unsettling new technology. Among their recommendations, the authors emphasize a need for the United States and its allies to develop a code of conduct for governments' use of deepfakes.

The research report, "Deepfakes and international conflict," was published this month by Brookings.

"The ease with which deepfakes can be developed for specific individuals and targets, as well as their rapid movement—most recently through a form of AI known as stable diffusion—point toward a world in which all states and nonstate actors will have the capacity to deploy deepfakes in their security and intelligence operations," the authors write. "Security officials and policymakers will need to prepare accordingly."

Northwestern co-authors include world-renowned AI and security expert V.S. Subrahmanian, the Walter P. Murphy Professor of Computer Science at Northwestern's McCormick School of Engineering and Buffett Faculty Fellow at the Buffett Institute of Global Affairs, and Chongyang Gao, a Ph.D. student in Subrahmanian's lab. Brookings Institute co-authors include Daniel L. Bynam and Chris Meserole.

Deepfakes require 'little difficulty'

Leader of the Northwestern Security and AI Lab, Subrahmanian and his student Gao previously developed TREAD (Terrorism Reduction with Artificial Intelligence Deepfakes), a new algorithm that researchers can

use to generate their own deepfake videos. By creating convincing deepfakes, researchers can better understand the technology within the context of security.

Using TREAD, Subrahmanian and his team created sample deepfake videos of deceased Islamic State terrorist Abu Mohammed al-Adnani. While the resulting video looks and sounds like al-Adnani—with highly realistic facial expressions and audio—he is actually speaking words by Syrian President Bashar al-Assad.

The researchers created the lifelike video within hours. The process was so straight-forward that Subrahmanian and his coauthors said militaries and security agencies should just assume that rivals are capable of generating deepfake videos of any official or leader within minutes.

"Anyone with a reasonable background in machine learning can—with some systematic work and the right hardware—generate deepfake videos at scale by building models similar to TREAD," the authors write. "The intelligence agencies of virtually any country, which certainly includes U.S. adversaries, can do so with little difficulty."

Avoiding 'cat-and-mouse games'

The authors believe that state and non-state actors will leverage deepfakes to strengthen ongoing disinformation efforts. Deepfakes could help fuel conflict by legitimizing war, sowing confusion, undermining popular support, polarizing societies, discrediting leaders and more. In the short-term, security and intelligence experts can counteract deepfakes by designing and training algorithms to identify potentially fake videos, images and audio. This approach, however, is unlikely to remain effective in the long term.

"Anyone with a reasonable background in machine learning can generate

deepfake videos at scale. The intelligence agencies of virtually any country can do so with little difficulty."

"The result will be a cat-and-mouse game similar to that seen with malware: When cybersecurity firms discover a new kind of malware and develop signatures to detect it, malware developers make 'tweaks' to evade the detector," the authors said. "The detect-evade-detect-evade cycle plays out over time…Eventually, we may reach an endpoint where detection becomes infeasible or too computationally intensive to carry out quickly and at scale."

For long-term strategies, the report's authors make several recommendations:

- Educate the general public to increase digital literacy and critical reasoning
- Develop systems capable of tracking the movement of digital assets by documenting each person or organization that handles the asset
- Encourage journalists and intelligence analysts to slow down and verify information before including it in published articles. "Similarly, journalists might emulate intelligence products that discuss 'confidence levels' with regard to judgments."
- Use information from separate sources, such as verification codes, to confirm legitimacy of digital assets

Above all, the authors argue that the government should enact policies that offer robust oversight and accountability mechanisms for governing the generation and distribution of deepfake content. If the United States or its allies want to "fight fire with fire" by creating their own deepfakes, then policies first need to be agreed upon and put in place. The authors say this could include establishing a "Deepfakes Equities Process," modeled after similar processes for cybersecurity.

"The decision to generate and use deepfakes should not be taken lightly and not without careful consideration of the trade-offs," the authors write. "The use of deepfakes, particularly designed to attack high-value targets in conflict settings, will affect a wide range of government offices and agencies. Each stakeholder should have the opportunity to offer input, as needed and as appropriate. Establishing such a broad-based, deliberative process is the best route to ensuring that democratic governments use deepfakes responsibly."

Provided by Northwestern University