The Rise of Artificial Intelligence and Deepfakes

AI-generated deepfake media is a growing threat to international security, yet deepfakes may also hold promise for counterterrorism. Through smart policies, public awareness campaigns and technical countermeasures, the threat of deepfakes may be mediated, while the promise harnessed responsibly.

DEEPFAKES THREATEN INTERNATIONAL SECURITY, HOLD POTENTIAL FOR ANTI-TERRORISM

The advancement of artificial intelligence (AI) is a growing concern for the international community, governments and the public, with significant implications for national security and cybersecurity. It also raises ethical questions related to surveillance and transparency. In a world rife with misinformation and mistrust, AI provides ever-more sophisticated means of convincing people of the veracity of false information that has the potential to lead to greater political tension, violence or even war.

Deepfakes—media content created by AI technologies that are generally meant to be deceptive—are a particularly significant and growing tool for misinformation and digital impersonation. Deepfakes are generated by machine-learning algorithms combined with facial-mapping software that can insert that data into digital content without permission. When execution is excellent, the result can be an extremely believable—but totally fabricated—a text, video or audio clip of a person doing or saying something that they did not.

Researchers have identified several possible use cases of deepfake technology with ramifications for the security sector, including the falsification of military orders to sow confusion among rank-and-file soldiers, discrediting political leaders and exploiting national tensions to promote polarization and discord. In June 2019, for example, a controversial video began circulating in Malaysia. The video appeared to depict Malaysia's Economic Affairs Minister engaged in physical relations with an aide. Despite the aide's attestation to the authenticity of the video, the minister and his political allies claimed the video had been doctored using deepfake technology, demonstrating the nefarious political potential of deepfake technology as a new form of political slander—or, as a new mechanism for obscuring truth.

One year later, in the spring of 2020, Extinction Rebellion Belgium released a fabricated video of then Belgian prime minister Sophie Wilmès appearing to connect the spread of COVID-19 to uncontrolled ecological crises. The group used footage taken from her recent address to the nation about the pandemic and generated a similar, fake speech with a script written by Extinction Rebellion. While transparently deepfake, this instance demonstrates how bad actors could use deepfake technology to promote the spread of misinformation.

Then, in March 2022, shortly after Russia began its invasion of Ukraine, the Ukrainian public was surprised to see a video of their

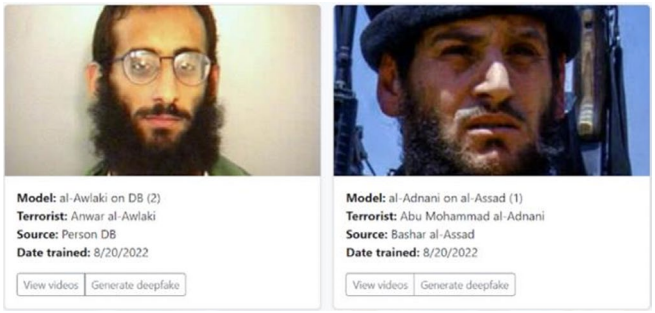
president, Volodymyr Zelenskyy, urging the military to lay down their weapons and surrender to the invading forces. As the video spread on social media and gained traction in the news, Zelenskyy's office quickly disavowed its authenticity. Indeed, the video had been generated using deepfake technology by Russian propagandists—the first high-profile example of a deepfake being weaponized in an armed conflict.

RESEARCHERS INVESTIGATE SOLUTIONS AND GUARDRAILS FOR CYBERDECEPTION

The Northwestern Security and AI Lab (NSAIL)—jointly housed by Northwestern University's Roberta Buffett Institute for Global Affairs and McCormick School of Engineering—is dedicated to performing fundamental research on AI technology, including deepfakes, that is relevant to questions related to cybersecurity and international security as well as when the use of AI or deepfakes is ethically or politically warranted.

NSAIL was launched in October 2022 as a partnership between the Northwestern Buffett Institute of Global Affairs and the Northwestern McCormick School of Engineering. Led by V.S. Subrahmanian, a Buffett Faculty Fellow at the Northwestern Buffett Institute and the Walter P. Murphy Professor of Computer Science, the new lab is currently working on over 20 distinct research projects related to AI, examining issues ranging from how to protect cities from drone attacks by terrorist organizations, to detecting deception in videos, to the implications of deepfakes for international conflicts.

The lab's Terrorism Reduction with AI Deepfakes (TREAD) project—developed by Northwestern University Ph.D. candidate Chongyang Gao, undergraduate Alex Feng and Subrahmanian—specifically investigates the implications of deepfakes for international conflict and terrorism mitigation while raising a central question for the global security community: Can deepfakes be used to counter terrorists and destabilize terror groups? NSAIL researchers have been at the leading edge of developing systems to generate realistic deepfake videos for countering terrorist groups, while also recommending extreme caution, the sparing use of deepfake technology and a deepfake code of conduct for governments.



Screenshot of TREAD used to put words in the mouth of two dead terrorists, Anwar al Awlaki and Mohammad al Adnani with trainers in English and Arabic, respectively.

EARLY FINDINGS AND RECOMMENDATIONS

NSAIL head V.S. Subrahmanian, in collaboration with Dan Byman of Georgetown University and Chris Meserole of the Brookings Institution, recently published a paper on the implications of deepfakes on international conflict, including the threats of "falsifying orders from military leaders, sowing confusion among the public and armed forces and lending legitimacy to wars and uprisings."

In addition to highlighting the ways in which deepfakes can be used to foster dissent, confusion and distrust among an adversary's public, military and media, the report also outlines a series of policy recommendations for liberal democracies balancing a vested interest in the accuracy of public information with strong incentives to deploy deepfakes against their adversaries, particularly in the context of armed conflict. Considering those incentives, Subrahmanian and his collaborators suggest that the U.S. and its democratic allies develop

a code of conduct for deepfake use by governments called a Deepfakes Equities Process based on the federal government's existing Vulnerabilities Equities Process, which guides decisions on whether newly discovered cybersecurity vulnerabilities are publicly disclosed or kept secret for offensive use against government adversaries. An "inclusive, deliberative process is the best way to ensure deepfakes are used responsibly," they wrote, and a Deepfakes Equities Process would determine when the benefits of leveraging deepfake technology against high-profile targets outweighs the risks "by incorporating the viewpoints of stakeholders across a wide range of government offices and agencies."'

A Deepfakes Equities Process would determine when the benefits of leveraging deepfake technology against high-profile targets outweighs the risks "by incorporating the viewpoints of stakeholders across a wide range of government offices and agencies."

DEVELOPMENTS TO WATCH

Deepfake technology has wide-ranging implications beyond international security, including the fabrication of criminal evidence, new forms of sexual harassment including deepfake pornography and/or privacy violations from employers trying to prevent deepfakes from happening in the first place. In the months and years ahead, governments, international organizations, institutions and businesses are likely to pay greater attention to the growing role of AI and the guardrails needed for risk mitigation through new and evolving technologies, policies and strategies. For example, in June 2023, the European Union took the first step towards regulating how businesses can use artificial intelligence. Undoubtedly, the EU will not be the only international body to take such steps in the near future and beyond.

If every piece of media content becomes suspect because of AI and deepfakes, Western democracies in particular will become ever more concerned about the potential erosion of trust needed for democracy to function. In response, expect an increase in the number of organizations, thinktanks and research organizations like NSAIL devoting time and resources to studying the evolving capabilities and threats of cyber deception, developing technical countermeasures to detect them and providing risk mitigation guidance and solutions.

NSAIL researchers and founder Subrahmanian will continue to advance the world's knowledge about AI technology across many domains, while presenting their findings globally. For example, in February 2023, the Dutch government hosted the first global Summit on Responsible Artificial Intelligence in the Military Domain (REAIM 2023). The summit convened stakeholders from around the world for dialogue on key opportunities, challenges, and risks associated with military applications of AI. Subrahmanian was an invited speaker, arguing for bringing insight on deepfake technology to the fore of current debates on responsible uses of AI within militaries.