Northwestern | McCORMICK SCHOOL OF ENGINEERING

Northwestern | BUFFETT INSTITUTE FOR GLOBAL AFFAIRS

Terrorism Reduction with AI Deepfakes

Deepfakes have proven divisive. While ordinary citizens enjoy the digital art created by products such as DALL-E, e.g., "a painting of a fox sitting in a field at sunrise in the style of Claude Monet," darker trends include the use of such techniques or by criminal groups for generating fake nude images of real persons by the Russian military for generating fake images of Ukrainian President Zelensky, among others.

NSAIL researchers have been at the leading edge of the development of systems to generate realistic deepfake video for countering terrorist groups. For instance, our prior work discovered that when there is internal dissension within the terror group Lashkar-e-Taiba (that carried out the 2008 Mumbai attacks), they carry out almost no terror attacks. A counter-terrorism strategy therefore might be to generate deepfake videos of LeT personnel criticizing others or trying to push agendas that they know are anathema to others in the group. The new TREAD system from NSAIL generates deepfake videos with realistic movement of facial muscles, eyelids, lips, and more, while preserving the audio characteristics of a terrorist.

Working jointly with researchers at the Brookings Institution and Georgetown University, NSAIL researchers have also proposed the development of a Deepfake Equity Program (or DEP) like the US Government's Vulnerability Equities Program for exploiting cyber-vulnerabilities in cyber-attacks for national security purposes.

What TREAD Does

TREAD develops deepfake videos of terrorists using a 2-phase process: training and operational use. Both phases involve inputs from two people – a "trainer" A who is typically an individual working for a security agency and a terrorist B. The greater the command that A has of B's native language, the better

During the *training* phase, TREAD requires a 15-20 minute video of the trainer A in a clean (noise-free, white background) environment. In addition, TREAD needs a 15-20 minute audio clip of the terrorist B as well as a good facial headshot of the terrorist. Using these three inputs and a suite of sophisticated deep learning based algorithms, TREAD learns a model *M(A,B)* that links the facial image and audio of the terrorist with the video of the trainer. In particular, *M(A,B)* learns how to transfer the eye/lip/facial muscle movements from the trainer's video onto the image of the terrorist in such a way that the speech of the terrorist is coordinated with such movements, while retaining the characteristics of the terrorist's own voice/audio signal.

In the *operational use* phase, the same trainer (A) records a new video *nv* with whatever the security agency would like B to say. The learned model *M(A,B)* is applied to the new video *nv* and a fake video *fv* is generated in which the terrorist says whatever is said in *nv*. Same words, but with the terrorist's voice, and muscle movements transferred seamlessly from the training video to the terrorist's face.

What TREAD Does

TREAD is capable of generating deepfake videos in any language (assuming of course that the selected trainer can speak the language), with both male and female trainers and terrorists (though the trainer should be of the same gender as the terrorist).

What TREAD Does Not Do

NSAIL researchers have not modified the metadata in videos generated by TREAD to avoid inconsistencies – as these kinds of edits are easy to incorporate. For instance, the metadata associated with a video might include latitude/longitudes of where the video was taken, the kind of camera used, and more. Such information is very helpful in identifying a video as fake. As an example, suppose a counter-terrorism agency creates a deepfake video of a terrorist based in the Sahara desert. If the GPS coordinates in the metadata suggest the video was shot near the Pentagon, the video would be quickly unmasked as fake. Most times, such metadata can be easily faked (e.g. by simple edits to the video file). We do not do this with TREAD's deepfakes which are only intended for demonstration purposes.

NSAIL researchers do not say what governments should or should not do or what TREAD should or should not be used for except to say that TREAD should not be used for any illegal or immoral purpose.

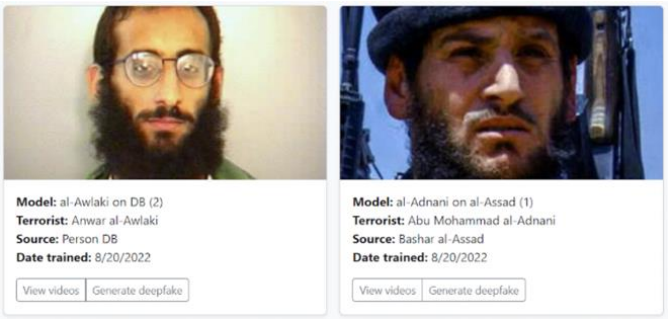

Fig 1. This shows a screenshot of TREAD used to put words in the mouth of two dead terrorists, Anwar al Awlaki and Mohammad al Adnani with trainers in English and Arabic, respectively.

However, we do recommend that governments publicly articulate a process that they will follow about when to use deepfakes and how to use them.

The Deepfake Equity Process

Jointly with researchers at the Brookings Institution and Georgetown University, TREAD researchers recommend that governments use deepfakes sparingly, even if the use is for valid counter-terrorism purposes. To determine whether a particular use of deepfakes is acceptable or not, we recommend that governments establish a "due process" for the use of deepfakes. This Deepfake Equity Process or DEP should:

- Be formulated after soliciting comments from the national security community
- Be published openly
- Be auditable by relevant government agencies in order to ensure that all operational use is compliant with DEP guidelines.

This does *not* require that governments disclose the deepfakes they have generated or the terrorists targeted using such deepfakes.

Additional Information

https://sites.northwestern.edu/nsail/projects/tread/

Video

https://sites.northwestern.edu/nsail/videos/tread/

PARTICIPANTS

Lead: V.S. Subrahmanian

Current: Alex Feng, Chongyang Gao

Brookings: Chris Meserole

Georgetown University: Daniel Byman

Northwestern | Security & AI Lab