TechTalk | Deepfake Technology
Saxon A.H. Knight, Head of Strategy & Government Partnerships, Reality Defender
25 October 2023
Summary Written by: Joseph Roskop, TLS Student Fellow

Purpose:

Cybersecurity and risk mitigation expert Saxon Knight participated in the TLS/PIJIP Tech Talk Series, where she discussed detection of AI-generated content and provided a framework for thinking through related issues.

Executive Summary

The need to quickly (in some cases—immediately) identify content as being inauthentic is only growing. One concern is the ability for misinformation and disinformation to go viral. AI-generated content can occur in any of four modalities: 1) video, 2) text, 3) audio, 4) images.

Ms. Knight drew a distinction between the ability to identify *Fake vs. Real* content, which can be established with some precision through technical means, as opposed to *True vs. False* content, which is a more challenging proposition.

Use of Deepfakes

Use of Deepfakes is generally on the rise due to the increased accessibility of consumer tools. Some AI voice-replicating software can accurately mimic a voice using only 10 seconds of audio data to train the model. There is particular concern around the use of Deepfaked content influencing political outcomes.

Industry Response

Reality Defender, where Ms. Knight works, relies on "Ensemble Models" to increase the efficacy of moderation tools. For example, a video with audio might yield different conclusions (authenticate or fake) depending on whether a video-specific or audio-specific model is used; therefore, using multiple models ensures greater reliability to under which components are likely AI-generated.

Future Goals

The hope is that one day these AI tools can be used to moderate millions of pieces of media content by running in the background similar to an antivirus software. Similarly, these tools present an opportunity to be proactive with one's cybersecurity posture rather than reactive. Regarding the future potential of these algorithms, Saxon Knight responded with, "The worst day [for a model] is today. It'll only get better, and it doesn't forget."