Oct 15 2019

Media AI Cybersecurity

Brief #3: The advance of deepfakes is spurring new countermeasures

12 min read

What's Happening

Deepfakes – AI-generated fake video, audio, and still images that seem real – have advanced rapidly since the early research projects and porn superimposed with celebrity faces uploaded to Reddit around 2016-2017. In Jul 2019, the CEO of image-verification startup Truepic suggested that "visually undetectable deepfakes" are less than 12 months away. Two months later – last month – deepfake pioneer Hao Li projected that **"perfectly real" deepfake videos were just 6-12 months away**. Li's accelerated timeline was partly driven by the Aug 2019 launch of wildly popular face-swapping Chinese app Zao, which lets users insert their face seamlessly into iconic movie/TV scenes. (A consumer uproar ensued when Zao's user agreement was found to be overly broad, raising the specter of identity theft and resulting in an inquiry by the Chinese government, blocking of Zao videos by WeChat, and ultimately changes to the user policy.) With the 2020 elections looming in the US, there is increasingly intensive pressure on industry from Congress and the public to figure out ways to counter the threat of deepfakes.

If industry projections prove out, we may have less than 12 months, and perhaps even less than 9, until perfect deepfakes are readily accessible.

Jump to **What It Means**

How deepfakes work

Currently, the most prevalent/effective methods for generating deepfakes are variants on a machine-learning approach called Generative Adversarial Networks (GANs). GANs, which have only been around since 2014, involve the interplay between two unsupervised machine-learning networks – a "generator" and a "discriminator" – working together in a feedback loop. The generator creates fake media from inputs (e.g. real images, videos, sound samples), and the discriminator tries to figure out which pieces of media in a dataset are fake. Results are sent back to the generator and discriminator for both to "learn" from. The two networks improve over thousands of iterations until the generator's output resembles something convincingly real. GANs have also been used for improving search and self-driving vehicles, with an extensive long tail of use cases.

The advance of deepfakes

According to a Sep 2019 report from Dutch deepfake detection company Deeptrace, **online deepfake videos doubled during the period of Dec 2018-Aug 2019** to 14,678. While 96% of these were pornographic videos placing celebrity or well-known women in fabricated situations, the rate of **growth and prevalence of deepfake tools and services are raising concerns about identity theft, fraud and misinformation**. In addition to open-source code available on platforms like GitHub for deepfake creation, there are now online forums offering downloadable tools with step-by-step tutorials, online deepfake services for as little as $2.99 per video, and deepfake creators offering videos for $30 and audio for $10 for every 50 words generated. There is even commercial voice-generating software now available (e.g. Lyrebird, iSpeech). **Deepfakes are growing in quality** (e.g. Zao's face-swapping app, full-body deepfakes that can mimic gait) **and are requiring less source data than before** (e.g. 3.7 seconds of audio, a few still shots or even just one). There are also emerging tools that make them easier to edit, using text-based video and audio editors.

While a lot of attention has been on political elections and foreign relations, **businesses are also susceptible to deepfake attacks**. There have been at least 3 deepfake-audio attacks on corporations this year, according to Symantec, one of which cost a firm $10M in funds that were wired to the culprits. In a separate case described by insurer Euler Hermes, the CEO of a UK firm thought he was receiving an order from the German CEO of the parent company to wire funds to a supplier in Hungary and complied, sending $243,000. Deepfakes can also be used to phish for personal or confidential information or other forms of social engineering – e.g. impersonating a Bloomberg reporter on social media to pump investors for information or your boss asking you to send a file. Biometric attacks on security systems that use voiceprint or facial authentication – which are used by a number of banks, credit card issuers, and crypto exchanges – are another risk.

Addressing the deepfake threat

There are a growing number of **tech firms, startups, and government organizations working on countermeasures** to address the deepfake threat:

Tech firms

- **Adobe** was early to this arena with Photoshop and later its Voco project in 2016 (a still-unreleased "Photoshop for voice"). It has been researching ways to detect image manipulation since 2016, with early efforts focused on watermarking. More recently, Adobe Research announced this year that it had begun working with collaborators at UC Berkeley (under DARPA's MediFor program) on a method to detect image manipulation with Photoshop. It will become part of a set of public forensic tools to detect the full gamut from amateur Photoshopped images to high-quality deepfake videos.
- In early Sep 2019, **Facebook and Microsoft teamed up with the Partnership on AI coalition and 7 universities to launch the Deepfake Detection Challenge**, which will run until spring of 2020. Participants will be given access to a database of realistic deepfake videos being developed by Facebook's AI researchers, which is expected to be released in Dec 2019. Facebook will also commit $10M to fund detection technology with grants and prizes.
- Google followed up a few weeks later by releasing a **dataset of 3,000+ deepfake videos** created using 28 paid actors. It is free for use by researchers working on detection tools. The dataset has been incorporated in the Technical University of Munich and University Federico II of Naples' **FaceForensics** project, which supports detection-tool development by offering a forensics dataset of nearly 1,000 YouTube videos manipulated using 4 different methods (500,000+ frames in total) and automated benchmarking of tools. In the same announcement, Google said that another **dataset of synthetic speech** – which had been shared with outside organizations for a fake-audio detection challenge – would now be freely available as well. Google previously developed voice-mimicking technology WaveNet in 2016.
- **Nuance Communications**, which sells voice authentication solutions to enterprise customers (incl. banks and card issuers), has been working on countermeasures for its own business since at least 2017. Its fraud detection uses personalized models of how a person speaks, types and behaves (e.g. how they use a mouse). It can also detect tiny skips where speech has been spliced together.
- **Symantec** researchers have also been working on technology to detect audio fakes. They have noted 4 areas of promise for future research: Face liveliness, contextual intelligence, texture investigation, and user interaction.

Startups

- **Forensic detection** is the most common approach to countering deepfakes, typically involving analysis of media using AI to identify whether it is fake – e.g. based on

discrepancies or inconsistencies. In addition to the tech firms, there are also smaller firms such as cyber risk company **ZeroFOX** (which opened up a deepfake defense toolkit for the security community), cybersecurity startup **Deeptrace**, video-authentication firm **Amber**, and voice-authentication firm **Pindrop** all advancing solutions. The challenge with forensic detection is that it will likely, by its nature, always be a step behind the deepfakes it trying to detect.

- **Digital trail** solutions help track the origin of media by verifying it the moment it is recorded using a digital watermark, fingerprint or "breadcrumb" (e.g. with data such as the device used, location, and timestamp). **Serelay** and **ProofMode by Guardian Project** are among the solutions in this space. There are also blockchain-based solutions for hashing/watermarking that prove a piece of media exists at a certain point in time, such as **Amber** (which does both forensic detection and digital trail), **Factom** (which the US DHS is testing to see if it can secure border surveillance cameras from tampering), **TFA Labs' Signed at Source** solution for IoT devices (which uses Factom), and **Truepic** (which has partnered with Qualcomm to integrate its tech into phones).
- **Social-network monitoring** identifies and remediates threats on social media accounts including deepfakes, taking action to prevent proliferation, and in some cases, helping take down accounts and even shape a communications response. Some players provide social-monitoring tools as part of a broader disinformation or digital-risk protection offering like Graphika, ZeroFOX and New Knowledge.

Government organizations

- The US Department of Defense's **DARPA** funded a **Media Forensics** (MediFor) program in 2016 that runs through 2020. MediFor has spent $68M+ producing some of the earliest tools for detecting deepfakes in digital imagery. It funded a deepfake detection challenge and awarded contracts to researchers – incl. SRI International, University of Amsterdam, and the Idiap Research Institute in Switzerland. Some of the techniques, however, used physiological signs (e.g. irregular or nonexistent blinking, strange head movements or eye color), sensor noise patterns, and/or statistical analyses which were useful in the early stages but will eventually be outmaneuvered.
- **DARPA** also this year launched a 4-year **Semantic Forensics** (SemaFor) project to detect what it calls "semantic errors" – inconsistencies in language, meaning or logic such as mismatched earrings in a photo. The project spans different types of multi-modal media with the intent of identifying and deterring disinformation (or "fake news") campaigns.
- On a related note, **California's governor just signed two laws banning deepfakes** of politicians within 60 days of an election and allowing civil lawsuits if an individual's image is used in sexually explicit material. Virginia also amended its "revenge porn" law to encompass deepfakes, and New York and Texas have proposed laws restricting deepfakes as well. At the federal level, **an array of bipartisan pieces of legislation was introduced over the past year** – criminalizing deepfakes (Malicious Deep Fake Prohibition Act of 2018), supporting research on countermeasures (IOGAN), directing the Department of Homeland Security to report on the risks of deepfakes (Deepfake Report Act of 2019), and

requiring disclosure of synthetic media (DEEPFAKES Accountability Act). Critics such as the Electronic Frontier Foundation, however, warn about the possibility of deepfake legislation hampering free expression and innovation.

There are also **researchers working on the leading edge of deepfake countermeasures**. One team of researchers from GSI Technology, Bloomberg, and University of Oregon is using animals – mice – to identify simulated voices by distinguishing between phonemes. Another team of researchers from New York University have demonstrated the possibility of adapting signal processors inside a camera to place watermarks on each photo. A USC group focused on inconsistencies through time (rather than by frame) – with specific attention on the movement of the eyes and mouth. Another USC-affiliated set of researchers in collaboration with others from UC Berkeley found useful fodder in facial expressions and head movements, with linguistic analysis to potentially be added later.

Other AI-generated fakery

While deepfakes have largely been used to describe fake video, audio and still images, there is a **parallel war underway in the realm of AI-generated text and fake news**. The OpenAI GPT-2 text generator can produce content realistic enough that OpenAI originally didn't release the full model for fear of abuse – such as deployment by malicious actors with armies of bots on social media. Other researchers have since followed, incl. Israeli AI21's HAIM as well as the University of Washington and Allen Institute for Artificial Intelligence's Grover. The Grover team is also working on the detection side of the war, along with DARPA's Semantics Forensics project (see above).

What It Means

It is still harder to create a convincing deepfake than it is to detect one. Today, detection tools can reasonably pick up deepfakes using combinations of statistical analyses. There are still not a lot of deepfakes being produced – relative to what is likely ahead – and the vast majority (96%) are in the realm of the pornographic rather than large-scale stock manipulation, geopolitical machination, or fraud. However, it appears that will change – and soon. **If industry projections prove out, we may have less than 12 months, and perhaps even less than 9, until perfect deepfakes are readily accessible.** While it might take some time for the technology advances to fully saturate the market, we shouldn't underestimate the rapid pace at which content can propagate across social media channels and at which technology can be distributed through cybercriminal communities. Some solution providers are already planning for a world in which perfect deepfakes are an everyday reality.

It is likely to be, as many have referred to it as, an arms race. We should expect periods where it is easier to create a perfectly real-seeming deepfake than it is to detect or counter it. At least one digital-forensic researcher is already saying they are currently outgunned – "The number of people working on the video-synthesis side, as opposed to the detector side, is 100 to 1." In the near term, the existing solutions will have the greatest impact **when we already suspect that a given piece of media is fake or in the context of a controlled pipeline where detection can be applied systematically** (e.g. voice-authentication vendors). Unfortunately, that means a lot of deepfakes will get through and the potential to do harm is substantial. On the other hand, **when a piece of media is particularly controversial, it is more likely that the state-of-the-art countermeasures will be applied**. Broad-based defense will require, at the macro level, a combination of technology, education of the public, and legislation. Appropriate legislation can help slow adoption so that countermeasures and public awareness have time to catch up. We may need to reconsider identity as a form of personal intellectual property in legislation, and institute laws that address new kinds of identity theft.

For businesses, audio is likely more of a near-term threat than video. The human vector in cybersecurity is typically the most significant source of vulnerability – as in the case of the fraudulent request to wire funds described above. Addressing the threat at an enterprise level will involve employee training as well as cybersecurity and technology solutions. Businesses should also look ahead as they plan to address the risk of deepfakes, considering the potential for all of the different types of harm they could inflict:

- **Fraud**: Use of deepfake audio or video to trick a business leader or employee into providing valuable proprietary information or money to a presumed-legitimate counterparty.
- **Market manipulation**: Release of synthetic media depicting, for example, a company CEO making an announcement about a company merger, financial results, or product release that influences perceptions in financial markets and moves the stock price.
- **Reputation sabotage**: Publication of a fake video, audio recording, or image of a business leader or employee in an illegal or embarrassing situation in order to influence public perception of the individual, company, and/or brand.
- **Extortion**: Use of fake media depicting a business leader in an embarrassing or illegal situation to demand money or favors under the threat of public release.
- **Systems infiltration**: Synthetic representations of people used to gain login credentials to proprietary systems through social engineering, or bypass voice/facial biometric security systems to gain access to sensitive data.

In the long run, we will likely gravitate towards **large-scale verification mechanisms**, where most media is verified and any unverified piece of media will be more open to question. In this hypothesized environment, all significant media content from a large company – e.g. CEO or executive interviews, TV appearances, earnings calls, and other communications – would be digitally "fingerprinted" or otherwise tracked for later verification. Corporate phone and video-calling systems would have built-in voice/video biometrics and fraud detection, with alerts to

notify parties of potential malicious activity. Consumer devices that generate media (e.g. phones) would also have similar features built in for stamping content.

If industry partnerships and ecosystem buy-in are required, it may take a long time (if ever) to achieve mass adoption of verification mechanisms. In countries with national facial-recognition programs like China (where it is an everyday facet of life used for payments and access to internet and mobile services), we will likely see much faster adoption of large-scale verification with backing from the government. On the other hand, some are saying that these solutions – if they become a de facto standard – may exclude people without access to verification technology.

The growing complexity of information warfare is reshaping our society – making people question the truth ("reality apathy"), enabling wrongdoers to deny the truth ("the liar's dividend"), influencing public opinion even when the content is known to be fake. We will have to rethink what constitutes evidence in court and how to assess reputation in the public sphere. Trust is the oil of our civil society and we are facing growing uncertainty around whom to trust. One critical danger is in gravitating towards systems that require a high level of administrative oversight and control, meaning less privacy and potentially less freedom.

With all of the negative possibilities of deepfakes overshadowing the news, **we shouldn't forget that video- and voice-synthesis technologies can be used for positive outcomes as well.**

- For one, deepfakes can be used to **anonymize people** – to give them privacy by generating a different face while keeping their expressions, personality, and body language.
- The ability for technology to speak with a human face and voice also has the potential to change the relationship between humans and technology – e.g. **virtual assistants or robots for the elderly or infirm**. It can help provide a human-like "companion," one that could even carry the voice of a loved one – addressing depression (a common comorbidity to chronic disease) in addition to providing reminders for medication adherence.
- In some cases, video and voice synthesis can help humans and technology **work better together through use of natural language** or by activating the human instinct for politeness. Google Duplex, for instance, can help conduct natural conversations to carry out useful tasks like making a reservation.
- Finally, as we've seen in Hollywood, these technologies can be used to **create and edit multi-modal content** – from live-action and animated characters to film-dubbing to podcasts to synthetic music. Will Smith's Junior character in the recent Gemini Man cost "tens of millions" to make. As deepfake tech advances, it has the potential to democratize CGI media production, lowering the cost of original content and opening up expansive avenues for entertainment, immersive mixed-reality experiences, and learning/training. We can expect more startups like Weta and Pinscreen to enter the field, along with more investment in adjacent technologies such as 3D motion animation and volumetric capture.

Disclosure: Contributors have investment interests in Microsoft. Google is a vendor of 6Pages.

Have a comment about this brief or a topic you'd like to see us cover? Send us a note at tips@6pages.com.