AI Insight Forum: Risk, Alignment, & Guarding Against Doomsday Scenarios

Statement of Vijay A. Balasubramaniyan
CEO and Cofounder
Pindrop
Dec 6, 2023

Majority Leader Schumer, Senators Rounds, Heinrich and Young:

Thank you for the opportunity to discuss the critical question of the threats that certain uses of Artificial Intelligence (AI) bring to bear and the kinds of defense mechanisms we will need to counter those threats.

The generative AI revolution is here. With its ability to augment human intelligence, create new revenue streams and provide great customer experiences there is a significant effort to commercialize these technologies in rapid fashion. Many of the companies that are innovating here at breakneck speed are becoming the fastest growing companies in history and the overall market is projected to be $1.3 trillion by 2032[1]. However, this rapidly advancing technology and its ability to mimic humans starts to create serious questions about whether any image, audio or video is real or an AI-generated deepfake. Among all the potential risks of generative AI, we believe deepfakes could cause the most significant harm as they have the potential to break all online trust.

My statement focuses on three areas of deepfakes. First, we look at how deepfakes undermine online trust and therefore simultaneously threaten and undermine online commerce, digital/social media, and communication. Second, we discuss a new class of technologies that are being used to combat deepfakes, ***liveness detection*** or the ability for systems to detect what is human or what is machine forensically. We set forth five tenets of a good deepfake detection technology and how liveness addresses these tenets. Third, given the focus by some on watermarking, we highlight the importance of layers of defense (known in security as defense in depth) and how both liveness and watermarking are needed to combat deepfakes. We also explain the importance of the continued use of biometrics and other factors to further illustrate the value of defense in depth. We conclude this note by providing five specific recommendations for Congress and policymakers to consider.

Pindrop's Expertise in Deepfake Detection

To provide context on our expertise in these areas, the company I started out of my PhD research at Georgia Tech, Pindrop, has been at the forefront of deepfake detection over the last 8 years,

[1] Bloomberg Intelligence: https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/

which enables us to provide you in-depth insights into these threats and the evolving defense against them.

Pindrop is consistently one of the top performers in the international ASVSpoof Challenge that has been held since 2015. This competition is organized by academic institutions and focuses on detecting manipulated speech, both synthetic and replayed. In this challenge, our technology was able to detect deepfakes with high accuracy levels (lowest error rate of 4.04% of a single system at the 2019 edition). To date, Pindrop's research has received about three hundred and fifty citations in deepfake detection publications (e.g. [1, 2, 3]). Our patented deepfake detection models are trained on a dataset of twenty million deepfake detection samples which, to the best of our knowledge, is the largest dataset in the industry. Pindrop has detected deepfakes with an accuracy of over 99% and detected content from previously unseen deepfake engines (zero-day deepfakes) with 90% accuracy.

In the field, Pindrop has been fortunate to learn from our real-time risk analysis for 8 of the largest banks, 5 of the top 7 insurance companies and some of the biggest retailers and healthcare providers to protect customer accounts from fraud and malicious takeovers. Our precise voice identification technology recognizes unique identifiers within the human voice that can enable its customers to prevent more fraud and deliver exceptional customer experiences. Pindrop has the largest database of known fraudsters that is built from intelligence acquired from each of our customers. This intelligence allows us to see what attackers are doing and the tools they are adopting to avoid detection. Since its inception, Pindrop has analyzed more than 5 billion voice calls, detected over 3 million fraud events, and saved organizations more than 2 billion dollars and counting.

The Problem: Deepfakes Undermine All Online Trust

The US public has already been targeted by deepfakes that spread misinformation related to war, elections, and to commit fraud. Deepfakes already exist in audio, video or image form and can be recorded or created live. A KPMG study reveals that the number of deepfake videos available online is increasing by 900% annually[2]. Pindrop's own research found out that more than 90% of US consumers have concerns about deepfakes[3].

Given their rapid proliferation, it is important to address this threat with urgency as deepfakes undermine all online trust:

1. Businesses increasingly distrust individuals when they are interacting with them. This undermines commerce.
2. Individuals increasingly distrust any information that they read, see or hear. This undermines news and social media.
3. Individuals increasingly distrust each other when they interact remotely. This undermines communication.

[2] https://kpmg.com/kpmg-us/content/dam/kpmg/pdf/2023/deepfakes-real-threat.pdf
[3] https://www.pindrop.com/blog/findings-in-our-deepfake-and-voice-clone-consumer-report

For example, with the latest technology it takes only 3 seconds of a person's voice to create a deepfake of an individual -- which is easily available in social media. As a result, the FTC has warned of the use of AI in parent/grandparent scams[4]. This attack has already been proliferating all through the US in states like Arizona, Illinois, New York, New Jersey, California, Washington, Florida, Texas, Ohio, Oregon and Virginia with families losing an average of $11,000[5]. Given the urgent need to address this threat, we next discuss liveness detection, one of the most promising technologies for combating deepfakes.

Liveness Detection and Important Tenets of Deepfake Detection

Liveness detection is a class of technologies that leverages attributes that come naturally to humans but are hard for machines to replicate at scale over sustained periods. This works on the basic premise that any deepfake generator creates artifacts and patterns which are distinctly different from natural human interactions. These patterns may not be detectable by humans but can be identified when analyzed by specially designed artificial intelligence tools. Examples of such patterns in audio include uniform pauses, spectral distortions, and the fingerprint of the synthetic vocoder. In video, examples of such patterns include blurry regions in the face and background, lighting/shadow anomalies and lip sync.

Liveness detection can help to restore trust back in our digital landscape. Organizations can take advantage of liveness detection within their account opening and transactional workflows to deny malicious actors the ability to use a deepfake to gain access to a user's account. Media organizations and fact-checking agencies can leverage this technology to enhance verification processes to bolster the trustworthiness of information. Social media platforms can give users the tools to verify the authenticity of content before sharing it with their networks and avoid the spread of misinformation. Communication systems can use this to ensure that individuals on each end are real humans and not AI.

For a deep fake detection system to be effective, it must have the following five tenets:

1. Being able to assess in ***real-time*** if the content is genuine or deepfake. This is particularly important for sensitive transactions (e.g., financial transactions) and information (e.g., political messages) where a few seconds delay is enough to cause substantial harm. Liveness detectors can analyze frames of audio or video. For example, a single second of audio in the lowest fidelity channel (call centers) has 8000 samples of a person's speech. This allows the number of anomalies to stack up rapidly and allow for real-time deepfake detection.
2. ***Continuously assess*** if the content is a deepfake in full or in parts. For instance, a manipulation of a few words in a political speech can change the intended message. Again, given liveness detection's ability to analyze frames of audio/video it automatically is able to determine whether segments are real or fake. A great example to see the need

[4] https://consumer.ftc.gov/consumer-alerts/2023/03/scammers-use-ai-enhance-their-family-emergency-schemes
[5] https://www.usatoday.com/story/tech/columnist/2023/05/16/artificial-intelligence-voice-phone-scams/70216185007/

for both real-time and continuous assessment of any audio/video content is the analysis of Senator Blumenthal's opening remarks at the Senate hearing on AI. The Senator began by speaking with his voice and eventually switched to a deepfake impersonation of his voice. In this video, providing a real-time score for every 3 second segment allows you to see the granularity of operation of an effective liveness detection system.

3. *Being **resilient** to* various types of degradations (e.g., noise, reverberation, channel compression and transcoding) and adversarial attacks (white-box or black-box attacks as described in a recent study conducted by the University of Waterloo). For further detail see Pindrop's liveness detection's robustness against the Waterloo attacks.
4. *Provide **Explainability*** in the system decision. This is particularly important in applications like forensic analysis and verification strategies that rely on multi-factor and flexible decisions where there is a trade-off between false acceptance and false rejection. For example, Pindrop's Liveness Detection provides a probability score and a reason code with every decision. This reason code can include the kind of deepfake being experienced and even the deepfake engine being used to create the content.
5. Finally, *being able to detect **zero-day attacks***, i.e., its ability to detect new deepfake engines that have previously never been seen. Deepfake engines are not one monolithic system and are created by a large set of components. Pindrop recently demonstrated the ability of liveness detection to identify previously unseen deepfakes with a very high level of accuracy.

Defense in Depth: Watermarking and Voice Biometrics Augment Liveness

In addition to deepfake detection, digital watermarking is also a viable path to distinguish between live and synthetically generated audio, video, or text. While deepfake detection is passive, watermarking is an active approach that requires the involvement of the content creator to insert a signature signal into the media. A good watermarking must be imperceptible, robust to various degradations and attacks, and have sufficient capacity to carry information about the owner, copyright, AI-generated flag, etc.

While watermarking is possible in some circumstances where the content creators are cooperative (e.g., Government entities and ethical industry players), we believe it is difficult to enforce it, particularly because of the availability of many open-source tools for GenAI creations that are used by both the public as well as bad actors. Additionally, watermarking technology has its own flaws as shown by a new study conducted by the University of Maryland in the image domain. In the phone channel, watermarking is prone to challenges such as limited channel bandwidth, speech degradation due to added noise and reverberation, transcoding, and packet loss. Hence, we strongly believe that the sole use of watermarking will not be enough, and it is preferable to combine this with other sophisticated tools for synthetic speech detection such as liveness.

Finally, there is a lot of talk about the death of biometrics due to deepfakes. However, liveness and biometrics determine two separate things. Liveness determines if a real human is part of an interaction or created a piece of content. Biometrics determines if the right human is part of an interaction or created a piece of content. A great example to see the power of biometrics is this

deepfake of President Obama[6]. While Obama's face was deepfaked, his voice was provided by a real human, in this case Jordan Peele. While liveness would detect that Obama's face was deepfaked, biometrics would determine that it is not Obama's voice since the anatomy of Jordan Peele's vocal tract is significantly different than that of Obama. Similarly, in the recent MGM hack that was the result of credential harvesting over the phone, a biometric defense would have protected the account. In security, the concept of using layers of defense to address different attack vectors is known as defense in depth. An organization's security posture therefore needs to build on top of existing tools and technologies by weaving liveness detection into an in-depth and multi-layered defense strategy.

Combining liveness detection which answers the question "Is it a real human?", with a comprehensive multi-factor fraud detection and authentication strategy that also answers the question "Is it the right human?" enables organizations to thwart attacks from fraudsters on multiple fronts. By leveraging all available risk signals on a live interaction (including voice, device info, carrier pathways, behavior patterns), organizations can weave a holistic picture of the identity of the person with whom organizations are interacting. This holistic picture of identity provides far greater assurance about the security of the organization and can further improve the effectiveness of liveness detection. Biometric tools such as voice, and facial recognition can continue to play a useful role as part of a multi-faceted, in-depth defense strategy.

Recommendations

I would like to conclude my comments, by submitting the following five recommendations for this esteemed group and for Congress and policymakers to consider, which we think will fortify the trust needed by our consumers and our businesses to operate as a society and for us to continue to adopt new innovations that drive economic growth and productivity.

1. Direct NIST to create liveness/deepfake testing and certification standards for organizations to ensure their tools are resilient to deepfake threats.
2. Establish a pilot program between industry and the federal government to tackle the proliferation of deepfakes through liveness detection capabilities.
3. Require social media platforms to disclose the use of Artificial Intelligence through indicator tags to educate users on which content is generated or manipulated by AI.
4. Explore ways to incentivize companies and carriers to offer consumers the use of voice biometrics and liveness detection tools that use voice-based verification, as part of a defense in depth strategy that leverages multiple factors for authentication.
5. Promote watermarking as a necessary capability for all content creators (both human or AI generated) but recognize that it is not sufficient alone to prevent and detect deepfakes.

[6] https://www.youtube.com/watch?v=cQ54GDm1eL0