Kluwer Arbitration Blog

Arbitration Tech Toolbox: Is Generative AI Now the Biggest Threat to Remote Hearings?

Sean McCarthy (ArbTech) · Saturday, April 13th, 2024 · ArbTech

Most of you will be familiar with ChatGPT and some will even be familiar with Midjourney and Stable Diffusion, but the recent launch of another type of Generative AI service has flown under the radar in the wider media as of now. However, it poses some of the most important practical and conceptual questions as to how deepfake video AI technology might affect the practice of international arbitration, in particular in the context of remote hearings today and into the future. Check out my video below to see this tech in action.

As briefly mentioned, the tool used in the video comes from a company called HeyGen, that can create what it terms 'Instant Avatars' from a user submitting just a 2-minute recording of themselves speaking to the camera. Recordings can be made using HeyGen's free trial and therefore the barrier to entry in using it is essentially non-existent, with the service's paid tiers allowing even more lifelike elements of the avatar to be introduced.

This is the new frontier of text-to-speech AI deep learning and deepfake technology. Using only a 2-minute recording, HeyGen's AI model can produce an incredibly lifelike video by recreating your facial expressions, mouth movements and body language, and synchronising it all together with speech following any script that you provide. The deep learning technology behind HeyGen and its competitors involves a set of competing algorithms in what is known as a Generative Adversarial Network ("GAN"). One algorithm is the generator of the 'fake' output, while the other is the discriminator, *i.e.* its function is to attempt to guess whether the output is fake or not. Thus, the generator algorithm's core function is to fool the discriminator, consequently increasing the likelihood of fooling the human viewer.

Some viewers of my video may have questioned its authenticity as 'me' from the beginning, specifically in relation to the strangeness (or lack) of one or two micro-expressions that humans can intuitively recognise in other humans. This phenomenon is known as the 'uncanny valley'—the sense of unease a person feels at the imperfect production of human likeness through the lack of certain lifelike details . But HeyGen also offers a premium service called 'Studio Avatar', whereby a subject is recorded in a studio setting and the avatar trained from this data faithfully reproduces the user's visual appearance, voice and accent. Further, HeyGen can now even translate what a person has said into multiple languages, all while matching the mouth movements to each translated language.

In my opinion, HeyGen often overcomes the uncanny valley issue. Up until now, even Hollywood, which invests substantially in this technology, has had a difficult time traversing the valley when digitally reproducing humans on-screen. A notable recent example of improvements in technology thanks to deep learning is the de-aged Harrison Ford in the movie, *Indiana Jones and the Dial of Destiny*, where Ford's character looks as he would have forty years ago for the first 25 minutes of the film. But Ford's de-aged self sometimes looks *too* perfect to be mistaken for actual footage of him from the 1980s.

Why HeyGen arguably produces more realistic avatars is through a mix of elements. Chief among them is through our collective conditioning to the video quality and environment of Zoom/Teams and other virtual video meeting software. The quality is often nowhere near as crisp as one might encounter in a movie or television show, the webcam camera angle is a fixed one (*i.e.*, the camera does not show us a more 3-dimensional view of the subject), and the background itself might be someone's home office or a virtual scene. This all plays into the hands of a GAN deep learning model by allowing those slight imperfections in the AI reproduction to go largely unnoticed by the viewer.

Actual Risks for Remote Arbitration Proceedings

Previous blog posts (here and here) in recent years have covered the plethora of different types of audio and video evidence that were and are at risk of falsification due to the sophistication of GAN-based AI models, and the potential ways to detect their use. Previous real-time methods of deep-faking used things like face-swapping (superimposing a recreated face on top of a human subject in a real-time video), which cannot withstand scrutiny due to the retention of individual characteristics (voice, speech patterns, body shape and body language etc.) of the person under the digital mask.

The most pertinent question then remains as to whether there are any current, practical risks to conducting arbitrations in a partially or fully remote environment. Many readers will of course realise that my video was not created in real-time, and therefore may assume that the technology is not yet able to do that, thereby relegating the dangers of real-time deepfakes to a conceptual rather than actual and present problem. But it is in the current context that possibly the most unsettling news for disputes practitioners comes into play. This video shows that HeyGen can already create real-time studio quality avatars, as well as ones that can reply to questions on the spot using whatever knowledge base the user may wish to train it on, from ChatGPT to any other type of custom large language model.

The principal risk with this type of tool is of course in relation to witness testimony over any kind of digital connection. If one were able to merely gain access to a video recording of a witness speaking and use that to create a completely controllable and realistic avatar that can be programmed to reply to questions in real-time based on customisable sources of information, then one of the foundations of the arbitral process, witness examination, becomes increasingly precarious. Even procedurally-focused remote conferencing finds itself on uncertain ground if an arbitrator or counsel could be at risk of having their likeness and voice falsified and reproduced at any moment.

The question posed in the title of this piece then comes into sharper focus. As with any wave of

technological development, there will be breakthroughs and then an ‘arms race’ to create tools that detect and combat misuse. While tools such as Intel’s FakeCatcher, and Sensity advertise the ability to detect deepfakes in real-time, where does that leave tribunals today in ensuring the integrity of the arbitral process without expending prohibitive party time and cost on new tools or experts to safeguard remote witness testimony? Are there more practical steps that arbitrators can immediately take to try and root out deepfake technology in witness testimony?

Detecting Deepfake Video Testimony in Practical Terms

As with a lot of technical fields, to be forewarned is to be forearmed. Since the advent in 2017 of deepfake content in popular culture, organisations around the world have taken action to educate society on detecting fake media of all kinds. Two prime examples of this are Australia’s eSafety Commissioner and MIT’s Detect Fakes initiative. Some academics have questioned whether teaching common strategies to recognise deepfake videos meaningfully improves detection rates. However, those studies did not focus on real-time video, where an interlocutor can directly attempt to push the limits of an AI model if he or she suspects manipulation.

Furthermore, real-time deepfake video is still not at a level of sophistication that one might encounter in a pre-recorded video. This is to do with the impossibility of post-processing or fine-tuning when presenting a ‘faked’ subject in real-time. Some of the core limits of real-time deepfakes such as HeyGen’s Streaming Avatar revolve around physical actions that are unexpected or complex, in particular body movement. For instance, an avatar currently cannot raise their hand or display a certain number of fingers on command, nor can they physically adjust their camera or seating position if requested to do so. As such, knowing that current deepfake development focuses almost entirely on facial accuracy rather than postural changes or body movements would allow a suspicious arbitrator or counsel to make simple, and otherwise uncontroversial requests of a witness in order to ascertain whether he or she is talking to a real person.

Updating current virtual hearing protocols may further assist in the practical detection of deepfake witness testimony, but the predictability of such guidelines could be used to create deepfakes that are designed to ‘beat’ these common methods. This again underlines the singular importance of practitioner education on the techniques by which AI models are trying to fool the human viewer and their current technological limits. Thus, while many of us may underestimate the current rate of improvement of synthetic media, this wave is about to hit the legal industry, and those unprepared to detect it will suffer the costs.

This entry was posted on Saturday, April 13th, 2024 at 9:27 am and is filed under Arbitration Tech Toolbox, Arbtech, Artificial Intelligence, Deepfakes, Technology
You can follow any responses to this entry through the Comments (RSS) feed. You can leave a response, or trackback from your own site.