Kluwer Arbitration Blog

Arbitration Tech Toolbox: Deepfakes and the Decline of Trust

Leonardo F. Souza-McMurtrie (University of Cambridge) · Wednesday, October 4th, 2023

An unsteady camera lens captures figures enshrouded in the flickering semi-darkness of the room. Men in suits whisper in Spanish over a heavy wooden table. The Equatorian judge, his face captured by the concealed camera, stands up, signalling acceptance of the proposition: his decision for a $3 million bribe.

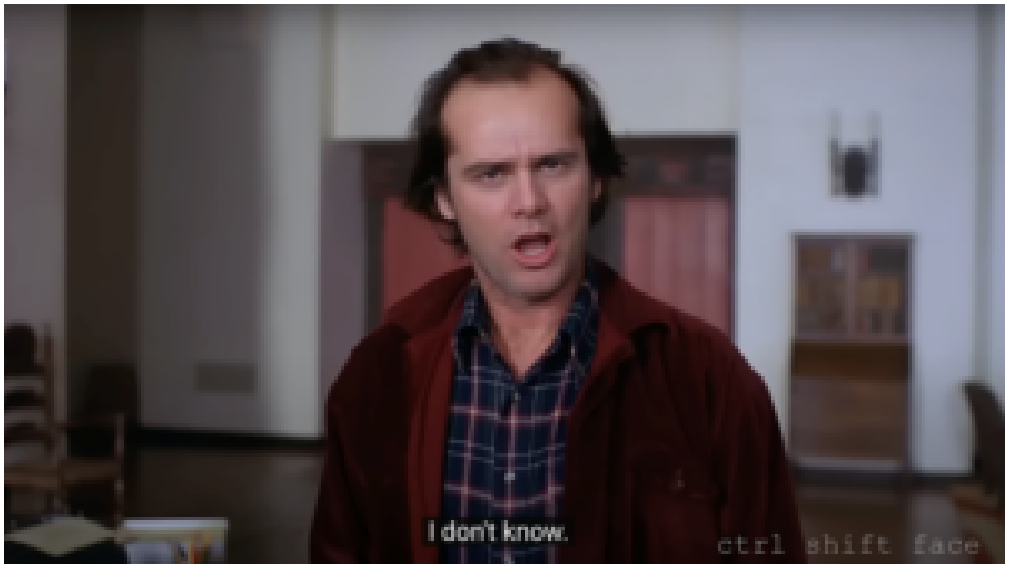
This scene was forever etched into the Lago Agrio lawsuit, an environmental battle fought across international arbitrations and courtrooms. For the arbitral tribunal in Chevron v Ecuador (II), this video was the smoking gun – compelling, irrefutable, and part of "*the most thorough documentary, video, and testimonial proof of fraud ever put before an arbitral tribunal*". Their position reflects the ones from the Tokio Tokeles award, in which video evidence gave "*an impression of the general level of confrontation*", and the Taftnet award deciding the "*number of people appearing in those recordings forcing their way into the premises, some in uniform, is credible evidence of a physical occupation*". This is the power of video evidence, the closest representation to the truth.

Until now, at least. As artificial intelligence rises, tribunals are facing a disconcerting reality: videos, photos, and audio recordings, once thought to be stalwart markers of authenticity, can be manipulated, distorted, or even created.

Deepfakes

Deepfakes are hyper-realistic simulations of video, audio, or image content, manufactured using machine learning algorithms capable of creating authentic-looking material. The lifeblood of deepfakes is a system known as Generative Adversarial Networks (GANs), which pits a pair of machine learning models against each other; one, the "generator", produces the fake, while the other, the "discriminator", assesses its authenticity. The two compete consistently learning and improving from each other's successes and failures, in a type of evolutionary arms-race.

GANs have become so adept that they can mimic a person's voice, gestures, and facial expressions to an uncanny degree of precision. Initially, this technology was used on celebrity faces and voices due to the surplus of data available, such as in videos matching Jim Carrey's face onto Jack Nicholson's body in a scene of The Shining:



But now, research has since evolved to create deepfakes based on a single photograph of a person's face, and several easily accessible websites give similar powers to the wide public:

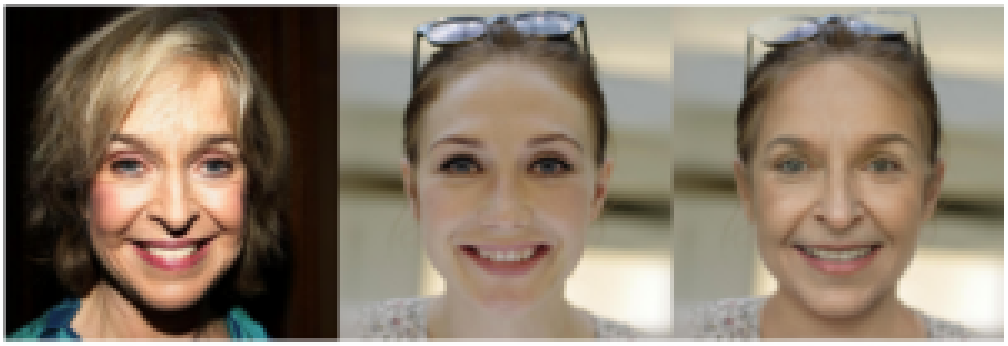


Figure 1. Example of a swapped face. Left: source image that represents the identity; Middle: target image that provides the attributes; Right: the swapped face image. All images are in 1024[2].

In 2018, US Congressmen even sent a letter to the National Director of Intelligence that "*deepfakes could undermine public trust in recorded images*", which ultimately became true when, in 2023, deepfakes became the interest for military psychological operations.

The Decline of Trust

The Decline of Trust in Evidence: Video, Audio and Photos

First, is audio-visual evidence still reliable? With the sophistication of deepfakes, distinguishing real from falsified evidence is challenging, as previously explored in this blog. A cleverly crafted deepfake video or audio clip could portray a party confessing guilt, making false statements, or engaging in other forms of misconduct – such as the one in the Lago Agrio arbitration. For instance, consider this custody battle in England, where a woman submitted audio evidence purporting her husband's abuse of their children, which upon expert scrutiny, was revealed to be a deepfake created with online tools.

Relying on falsified evidence can distort the outcome of arbitrations and, at worst, lead to miscarriages of justice. This can also impact time and costs if tribunals require experts to verify the legitimacy of audio or video evidence.

The Decline of Trust in Counsel: The Deepfake Defence

Conversely, the mistrust in certain kinds of evidence will generate what Rebecca Delfino called "The Deepfake Defence". In this strategy, lawyers capitalise on the scepticism about evidence to sow doubt about its legitimacy, even when it is indeed authentic. This creates an environment where counsel can exploit the arbitrators' apprehension over potentially counterfeit evidence, challenging the credibility of any piece of evidence at any time. This tactic has been inaugurated in the trials involving the invasion of the US Capitol. In arbitration, tribunals will have to decide whether to trust the counsel and consider the defence, determining the production of expert evidence and delaying the proceedings, or to disregard the defence *prima facie*, on the risk of undermining their awards afterward. Traditional guidelines, such as the CIArb's Code of Conduct or the CIArb's Guideline on the Use of Technology, predate the deepfakes and do not deal with this tactic, which lies at the intersection of technology, law, and ethics, masking itself as a genuine defence. There is nothing on how tribunals can avoid the conundrum, which can only make it worse.

The Decline of Trust in Hearings: Videoconferencing

The third issue, the most disconcerting one, involves videoconferencing and the collection of oral testimony.

Deepfakes compromise the authenticity of remote interactions during arbitrations. A deepfake model may be used to impersonate a witness, creating the illusion that the correct person is providing testimony when, in fact, they are not. A model can convincingly mimic voice, tone, facial expressions, and even mannerisms, making it incredibly challenging to ascertain whether the person seen and heard on the screen truly is who they claim to be. Even within specialised sectors, immediate verification of the truth remains impossible. For instance, deepfakes have successfully breached bank security systems and duped financial officers into transferring funds to fraudsters.

Disturbingly, arbitrators themselves can be "deepfaked" into video hearings. The forgery can be as simple as opening the eyes of a sleepy or inattentive arbitrator, or as severe as superimposing an arbitrator's face and voice onto someone else's body, forging the arbitrator's virtual participation. While this might seem far-fetched, parties made similar allegations regarding arbitrators' *written participation* in their awards. Indeed, accusations that arbitrators did not pen their awards, backed by stylometric statistical models, have led to annulment proceedings before.

The Decline of Trust in Awards: Annulment Proceedings

Deepfakes are a threat to the credibility of arbitral awards. Consider the implications for annulment procedures where the authenticity of witness testimonies, obtained during a video hearing, is questioned. Similarly, doubts can arise over the participation of an arbitrator in meetings, evidentiary hearings, or even intra-tribunal deliberations. Courts may find themselves wrestling with the uncertainty surrounding whether specific witnesses were heard, and if so, by whom. Without security measures, any video or audio interactions among the counsel, tribunal, experts, and witnesses can be regarded as untrustworthy.

Potential Preemptive Measures

Detecting Deepfakes with Artificial Intelligence

Ironically, one of the most promising solutions to the deepfake problem also lies in artificial intelligence. Just as AI creates deepfakes, it can detect them. Researchers are designing AI to

identify deepfakes by looking for subtle inconsistencies overlooked by the human eye, such as incoherent facial expressions or signs of blood flow on the skin of video participants.

Watermarking for Witness Evidence Collected via Videoconferencing

Recorded hearings can be watermarked to safeguard them from malicious subsequent alteration. Digital watermarking is a promising measure to counteract Deepfakes. The technique, originally used to protect copyrights, involves the embedding of a unique set of data, or a 'watermark', into digital media. The watermark, intertwined with facial identity features, becomes sensitive to face swap translations (i.e., Deepfake), but robust against conventional image modifications such as resizing and compression.

Specialised Videoconferencing Software

The solutions above cannot prevent participants from "deepfaking" their participation during remote conferencing. And usual videoconferencing software has no means to guarantee that no deepfake model is being used. Counteracting deepfakes requires specialised videoconferencing software capable of controlling participant machines to ensure that no deepfake models are running. Analogous systems have been implemented by testing services like ETS or Pearson for remote test takers.

Such software should have continuous monitoring to detect unauthorised programs running on the participant's device and will need to operate within a secure and encrypted environment, to safeguard against breaches, which will certainly raise valid privacy concerns in its trade-off between privacy and cybersecurity.

What can we, Lawyers, do?

The answer is – at best – unclear. I suggest three options below. But what is abundantly clear is that ignoring the issue or refusing to recognise digital threats is akin to blinding oneself and risking the integrity of real arbitration cases.

Criminal Reporting: Using forged evidence, or fraudulently representing oneself in judicial matters, is not merely an ethical breach—it is a crime in most jurisdictions, carrying severe criminal liabilities. Arbitrators must remain vigilant, not only to challenge such uses but also to report illicit activity to the authorities.

Training: While it is true that some state-of-the-art deepfakes can challenge detection, a reasonably trained eye can still discern the majority of live deepfakes, especially during real-time

events like hearings. Training on the latest deepfake creation and detection methods can arm practitioners with the skills to identify manipulations, or at least sense when something is wrong.

Establishing Protocols: We should think about standard protocols for authenticating digital evidence and communications. These can include cross-referencing the metadata, seeking expert testimony, or implementing the watermarking or specialised videoconferencing techniques outlined above, all of which will require both openness and specialisation from our industry members.

Conclusion: Simulacra and Arbitration

Arbitration is caught in Baudrillard's hyperreal, where the mirage is indistinguishable from the oasis. Witness evidence, counsel, hearings, and even arbitral awards stand on shifting sand. While preemptive measures – detecting deepfakes with AI, digital watermarking, and specialised videoconferencing software – stave off the encroaching chaos, they offer an incomplete solace: they are limited in scope and do little to deal with the decline of trust in evidence, for instance. Worse, their fight is akin to a game of Whac-A-Mole, with each victory only leading to a new, more sophisticated challenge. Why? Because deepfakes are fundamentally evolutionary; they adapt, learn, and grow, perpetually continuing the struggle.

The urgency cannot be overstated: deepfakes are not a terror of the future. Tribunals must reckon with this predicament today – during videoconferencing, testimony collection, and deliberations. Regardless of one's inclination to recognise the situation, tribunals already wander the *desert of the real*, where trust is an alarmingly scarce resource. Some may hope to sidestep this issue, but the choice to remain unguarded may have consequences. And those who prepare will, perhaps, face the most pressing issue of all: how does one parse the real from the simulation when deceit wears the skin of authenticity?

Further posts in our Arbitration Tech Toolbox series can be found here.

This entry was posted on Wednesday, October 4th, 2023 at 8:54 am and is filed under Arbitration Tech Toolbox, Artificial Intelligence, Deepfakes
You can follow any responses to this entry through the Comments (RSS) feed. You can leave a response, or trackback from your own site.