The Paradox of Artificial Intelligence in the Legal Industry: Both Treasure Trove and Trojan Horse? – The Perils of Deepfakes

Kluwer Arbitration Blog
July 18, 2021

Ingrid A. Müller

Please refer to this post as: Ingrid A. Müller, 'The Paradox of Artificial Intelligence in the Legal Industry: Both Treasure Trove and Trojan Horse? – The Perils of Deepfakes', Kluwer Arbitration Blog, July 18 2021, http://arbitrationblog.kluwerarbitration.com/2021/07/18/the-paradox-of-artificial-intelligence-in-the-legal-industry-both-treasure-trove-and-trojan-horse-the-perils-of-deepfakes/

The legal industry has benefited tremendously from recent technological advancements, leading to the expansion of Legal Tech as the driving force for progress in this field. More and more tools – more or less Artificial Intelligence ("AI")[fn]Generally, the term "Artificial Intelligence (AI)" refers to machines capable of replicating human intelligence. However, the technological status quo is rather limiting, with mainly so called "weak (or narrow) AI", *i.e.*, machines able to replicate (and outperform humans at) specific tasks, being available. In contrast, "strong AI", *i.e.,* machines able fully simulate the human mind, are (still) a log way out.[/fn] reliant – are developed to successfully simplify, automate, and expedite the work of legal professionals. To name just a few: contract automation services (e.g., Lawlift), e-discovery software (e.g., Everlaw), case management applications (e.g., App4Legal), information aggregators allowing for more informed decisions in choosing ADR neutrals (e.g., Arbitrator Intelligence), or litigation prediction solutions (e.g., Lex Machina).

However, until recently the legal industry was reluctant to fully embrace

technology, despite growing interest. Notwithstanding the many setbacks and challenges posed, it was the current pandemic that acted as the accelerator for the largescale acceptance of technology in this field, though some are still skeptical.

The international arbitration community was particularly quick to adapt to the new socio-economic reality with the help of technology, owing to the inherent flexibility of this dispute resolution method. From the very beginning, most arbitral institutions actively employed mitigating measures by adopting new procedures and issuing guidelines to encourage virtual hearings, while some arbitral institutions even drafted protocols on the conduct of such virtual hearings. In doing so, they addressed several potential perils, like hacking – which turned out to be a big problem – but some issues remained (e.g. specific due process concerns) so the actual measures implemented depended on the experience and tech savviness of the participants.

At a slower pace and with outright unwillingness in some cases even the judicial system eventually employed largescale virtual video and/or audio hearings and adopted relevant rules.

Yet, despite the wide implementation of officially sanctioned remote means of communication, the danger posed by potentially unethical technological advancements, such as AI manipulated media, was largely disregarded.

The problem with AI-generated or manipulated media

Nowadays, anyone can pretend they're someone else in the online realm via a misrepresented photo, video or audio recording, and lately apparently even in live transmissions, most of the time with the help of a simple app on a smartphone without necessarily having any tech expertise. Technologies once extremely expensive and only used by experts are now imbedded in virtually every electronic camera. Novel audio-video editing software can create forged audios and/or videos of anyone and have them say or do potentially anything.

This kind of manipulated media is colloquially called a *Deepfake.* It's a synthetic or altered media based on "deep learning", itself a subfield of machine learning inspired by the human brain and employing huge sets of data with the help of AI.[fn] *Deepfakes* are sometime distinguished from other manipulated media, like

Cheepfakes or *Shallowfakes* (which are of a lower quality, created with simpler/cheaper tools, and less AI reliant). In this paper, however, all manipulated media will be labeled "Deepfake" since this will arguably be the norm as the technology advances.[/fn] The concept has long been used for special effects in the movie industry but only recently similar software became available to the public at large, getting more and more advanced each year. Some are going as far as to consider deepfakes the "future of content creation" based on recent reports of news anchors being replaced by deepfake versions of themselves.

And this is just the beginning – AI can now generate fake people virtually indistinguishable from real ones, famous paintings are coming to life, hologram concerts have been employed for years, and soon enough we'll have multisensorial interactions with our long past ancestors. Analogous manipulation is possible beyond media presenting people – for example, it seems geography as we know it is in danger as well with deepfakes potentially becoming a security threat and posing challenges for geospatial agencies and the entire intelligence community. For that matter, for the legal community as well.

So, even if there are many positive uses for deepfakes, the problem is that the technology has become so advanced that we're almost no longer able to rely on our own senses in distinguishing fake from real. We can't unquestionably believe anymore what we see and hear even in direct interactions. Principles by which we're normally abiding in uncertain situations – like "trust your own eyes" or "a picture is worth a thousand words" – are becoming obsolete. There's no illusionist to watch out for in real time, the "magic"[fn]*"Any sufficiently advanced technology is indistinguishable from magic"*, Arthur C. Clark.[/fn] is happening before the "act" is even presented, and the illusion is "real"[fn]If we see and hear something with our own eyes and ears, can we say it's not real? Based on what criteria would *real* be defined objectively since it relates to a subjective experience, purpose, or perspective? To that end, (digital) reality can no longer be unqualified…[/fn]. Objective truth is getting increasingly difficult to ascertain,[fn]Since fake-detection measures are always immediately countered with newer technologies, paradoxically, the more we'll try to determine the genuineness of digital media the less possible it will be. So — analogizing this with the wonderfully weird world of quantum mechanics — there's a Heisenberg (like) uncertainty to it in the long run.[/fn] bringing epistemological relativism to new heights[fn]Facts themselves become relative. Take, for example, the recording of someone with the exact

features of person X (face shape, eye and hair color, voice, etc.) of such high technical quality that for all intents and purposes it seems a genuine recording of person X. Would it be a *true* or *false* to say that it *is* a recording *of* person X? A random viewer would *know* the (subjective) truth to be what he or she is able to asses through their own senses (i.e., *true*), while the person apparently in the video or the one who manipulated the media would *know* a different (subjective) truth (i.e., *false*).[/fn] and putting Schrödinger's cat to shame.

Granted, there's a trove of software that can be used to detect manipulated media through diverse methods like heartbeat detection, eye reflection mapping, or lip-sync analysis as most deepfakes are not (yet) very sophisticated. However, not only that are not infallible but fake detection counter-measures will always be one step behind as the technology progresses exponentially. For all intents and purposes, this is an "arms race". A digital one.

Thus, such technological advancements pose an excessive risk of unethical use and are potentially threatening the authenticity of the online identities of the participants to and of the evidence presented in (virtual) legal proceedings, with huge implications on the safety and security of the proceedings, on due process, and on the overall legal certainty of the outcome raising many challenges for the justice system as a whole.

Although arbitration seems especially vulnerable to the dangers of Deepfakes[fn]Even if experts could determine the authenticity of documentary evidence in advance, the identity of the participants in virtual hearings would still have to be verified in real time which would raise additional impediments (legal and otherwise), *e.g.*, higher costs or privacy issues.[/fn] since it's more difficult to implement adequate fake-detection measures in private settings, the potential implications are too serious for the entire legal system for this not to be addressed by all important actors.

The solution(s)

Generally, best suited to implement enforceable preventive measures are the governments. Some are already researching ways to counteract the dangers of deepfakes with enticing stratagems for scouting the best proposals, like prize competitions, but so far there are no coordinated global efforts and no cohesive

policies, not even at the national level.

Second in line are social media companies, in a position to enforce virtually any "behavior – modulating" terms of use throughout their platforms.[fn]Whether or not such enforced behavior is infringing on fundamental rights, like free speech, as well social media's potential liability for their users' behavior is debatable[/fn] But, for the most part, the range of restrictions employed by them is too narrow, thus inefficient.

Additionally, other private and public actors are working on identifying potential solutions and coming up with diverse ideas, like insurance coverage or adding some type of graphic label to the manipulated media. Still, one would have to first identify the media as deepfake, which, as we've seen, is getting harder to do.

A solution more apt to resolve the problem from the "digitally inceptionist" perspective would be to create an origination label, *i.e.,* embedding digital "fingerprints" in the relevant media, by capitalizing on the emergence of blockchain technology. The art world has already adopted equivalent procedures by ingeniously and very lucratively using so called NFT's (non-fungible tokens) to certify the uniqueness of digitally stored art. Similar proposals are also being tested for legal purposes by governmental institutions.

However, no matter what solutions are ultimately adopted (even if sanctioned by legislative bodies) the justice system must scrutinize them first in order to be widely accepted at the societal level when it comes to legally bounding issues.

As for the legal *status quo*, things are debatable. Anything from harassment laws to copyright laws to privacy laws to consumer protection laws could apply, as appropriate. Nonetheless, there are inherent limitations in solving novel problems with old tools. For example, establishing causality to discern who would be liable for an incident caused by an Autonomous Vehicle – the manufacturer, the software developer, the user, or the AI? This last option has potentially controversial AI legal personhood implications but was suggested by some for the purpose of insurance coverage. In any case, existing norms may prove anachronistic.

So, an important part of any successful strategy would be to ensure the appropriate legal framework. On this front, no comprehensive steps have been taken yet but recent EU and US (proposed) legislation is promising.

In conclusion, the best approach would be to involve all responsible factors from the beginning, with a multi-disciplinary approach. This would allow for quicker identifications of threats and improved decision making and implementation strategies, rendering the best results.

The Moral of the Story?

It's only fair to *question AI even though it comes bearing gifts.*[fn] Paraphrase of the old Latin saying of "Timeo Danaos et dona ferentes" (I fear the Greeks even when they bear gifts), spoken by Laocoön as a warning on potential dangers posed by the Trojan Horse. According to Vergil's Aeneid (II, 49), see Christopher Francese and Meghan Reedy (eds.), Dickinson College Commentaries on Book 2 of the Aeneid.[/fn]

Although Tech innovations should be welcomed as agents of progress for the legal profession, to ensure the future integrity of the justice system, we have to prevent AI innovations with a high risk for unethical use, like *Deepfakes*, from becoming justice's nemesis. And – no matter how implausible they may seem today – we can't ignore theories contending that to maintain the relevance of human lawyering it's better to have technology as a mere enabler instead of the driving force for innovation. So, while we're enjoying the obvious benefits of AI, let's not lose sight of the potential perils.

The sooner we act, the easier it will be to implement the necessary checks and balances. To develop digital forensics as an interdisciplinary field to easily recognize and address such dangers. To include the necessary restraints in the core structures of AI. To envision and agree on legally relevant AI ethical principles and on implementing bias-free procedures.

Si vis iustitia, cole aequitas.[fn]If you desire justice, cultivate fairness (equity). Play on words based on the motto of the International Labor Organization: "Si vis pacem, cole iustitia" (If you desire peace, cultivate justice). Even if justice and equity are sometimes used interchangeably, the two are sensibly different.[/fn] *Si vis [AI] aequitas, para pactum*.[fn] If you desire [AI] fairness, agree on it. Andrew Carnegie, promoter of universal peace, reportedly said during the National Arbitration and Peace Congress of 1907: "Si vis pacem, para pactum" (If you desire peace, agree to keep it) itself a paraphrase of the Latin adagio of "Si vis pacem,

para bellum” (If you desire peace, prepare for war) often attributed to Publius Flavius Vegetius Renatus and to this day motto of the UK naval warfare force, the Royal Navy. Carnegie was advocating for a safer way to keep peace, in opposition with the time’s usage of deterring enemies by show of force. See “The National Arbitration and Peace Congress at New York.” The American Journal of International Law, vol. 1, no. 3, 1907, JSTOR. The idea is that in order to ensure AI fairness we should first agree that we need it, how to define it, and how to implement it and keep it.[/fn]