Is That You? Deep Dive Into Deepfakes

Authors: Dean Gerakiteys (Partner) Clayton Utz, **Lex Burke** (FTS Director), Clayton Utz and **Natalie Coulton** (Lawyer), Clayton Utz (and now at Paul Hastings UK)

*While they may seem harmless and just a bit of fun, the artificial intelligence (**AI**) technology behind deepfakes also lends itself to more nefarious applications, and has the potential to expose both businesses and individuals.*

Having retired from acting in March last year, Bruce Willis' appearance in a recent advertisement for Russian telecommunications operator Megafon quickly gained interest from worldwide media. Even more surprising were the headlines which seemed to indicate Willis had in fact sold the rights to his face to a deepfake company called Deepcake, which had utilised artificial (AI) and machine learning to transform a Russian actor into an uncanny digital twin of the Hollywood heavyweight.

With deepfakes gaining increased attention, both individuals and businesses are turning to the cutting-edge technology to both explore the possibilities and understand its limitations. Mainstream access to the complex algorithms and software which facilitates the creation of believable lookalikes is growing, as is the ability to generate hyper-realistic and almost undetectable synthetic media.

We'll be taking a deep dive into the world of deepfakes, covering some of the key issues which deepfakes raise, both from a branding and intellectual property point of view as well as exploring some of the broader legal implications of this exciting new technology:

- what deepfakes are and how they can be used;
- the legal implications which attach to deepfakes and what the current state of the legal landscape looks like, both here and overseas; and
- some recent examples of deep trouble with deepfakes (including Bruce Willis' Russian digital twin) and some practical tips for recognising, deploying or protecting against deepfakes.

To kick off, we will begin with some background on deepfakes.

Deepcake deepfake

For those not already familiar with the technology, a deepfake (a portmanteau for a deep learning fake) is a form of digital media – a photo, video or sound file – that has been manipulated to generate a hyper realistic but false depiction of a person or thing doing or saying something they have not done or said. Despite being around since about 2017, deepfakes have only recently grown in popularity as technological advances have made it possible to generate higher quality and more convincing dupes of ordinary people or objects, as well as making the technology more accessible.

Deepfakes are produced using AI software which draws on a large information base of photos or recordings of the person to generate the deepfake content, with no need for the information base content to be related to the target clip. The generation of a deepfake requires two types of machine learning or AI technology:

- an autoencoder reviews the information base images or videos to understand how a person or object looks from various angles, in various light settings and in various moods; and
- Generative Adversarial Networks (**GANs**) then detect and improve flaws in the deepfake with multiple rounds, making the deepfake more and more believable.

This means that home videos or interview snippets could be used to generate a deepfake of an entire Hollywood movie scene. To generate Bruce Willis' digital twin in the deepfake advertisement, Deepcake used 34,000 images of the actor, taken largely from his appearances in Die Hard and The Fifth Element, giving the copycat Willis a much younger appearance, similar to his 1980s and 1990s self.

Uses (and misuses) of deepfakes

To date, deepfakes have already been used in a number of humorous ways, including giving famous paintings new facial expressions (as seen in the images below), as well as allowing members of the general public to superimpose their own face onto an iconic Spiderman scene. Another popular use for the technology has been to recast movie scenes with a fan favourite actor or actress who ultimately didn't take the role. While most of the mainstream uses for the technology have involved users swapping their own face for that of a celebrity or creating a video of a celebrity saying or doing something they have not done or said, the technology is also being explored for other exciting applications, such as for detecting tumours.

Unfortunately, the nature of this technology also lends itself to more nefarious applications, with deepfakes also blamed for the production of false political statements, conspiracies about the failing or recovering health of key public figures and material intended to greatly damage the reputation of celebrities and world leaders. The Australian eSafety Commissioner has also warned that deepfakes could be used to generate fake news and consumer confusion as well as potentially to intimidate individuals or companies, with the potential to cause financial loss, reputational damage and compromised consumer loyalty or reduced confidence in brands or companies (including from shareholders).

There is even the application of deepfakes in warfare; the Ukrainian Government is suspected of hacking a Russian television channel to broadcast a deepfake President Putin declaring mobilisation and martial law.

Real (deep)fake legislation: other jurisdictions' attempts so far

Australia has no specific legislation which addresses the potential misuse of deepfake technology, and we have yet to see a case concerning a deepfake reach the Australian judicial system. Other jurisdictions, however, have begun the process of legislating to address the potential for deepfakes to be misused.

Due to the growing number of deepfakes and their potential harmful uses, including with respect to politicians, the US states of California, Texas and Virginia have attempted to pass legislation regarding deepfakes. Former President Trump also passed national legislation in the US which required reporting on the foreign weaponization of deepfakes, including those targeted at disseminating disinformation relating to US elections and established a "Deepfakes Prize" competition to encourage research or commercialisation of deepfake detection technologies.

Both China and the European Union have also taken positive steps to combat harmful deepfakes, although China has also experimented with using deepfake news anchors to deliver news broadcasts.

A related issue that pervades the regulation of deepfakes is the myriad of practical difficulties with enforcing any legislative restrictions, particularly as:

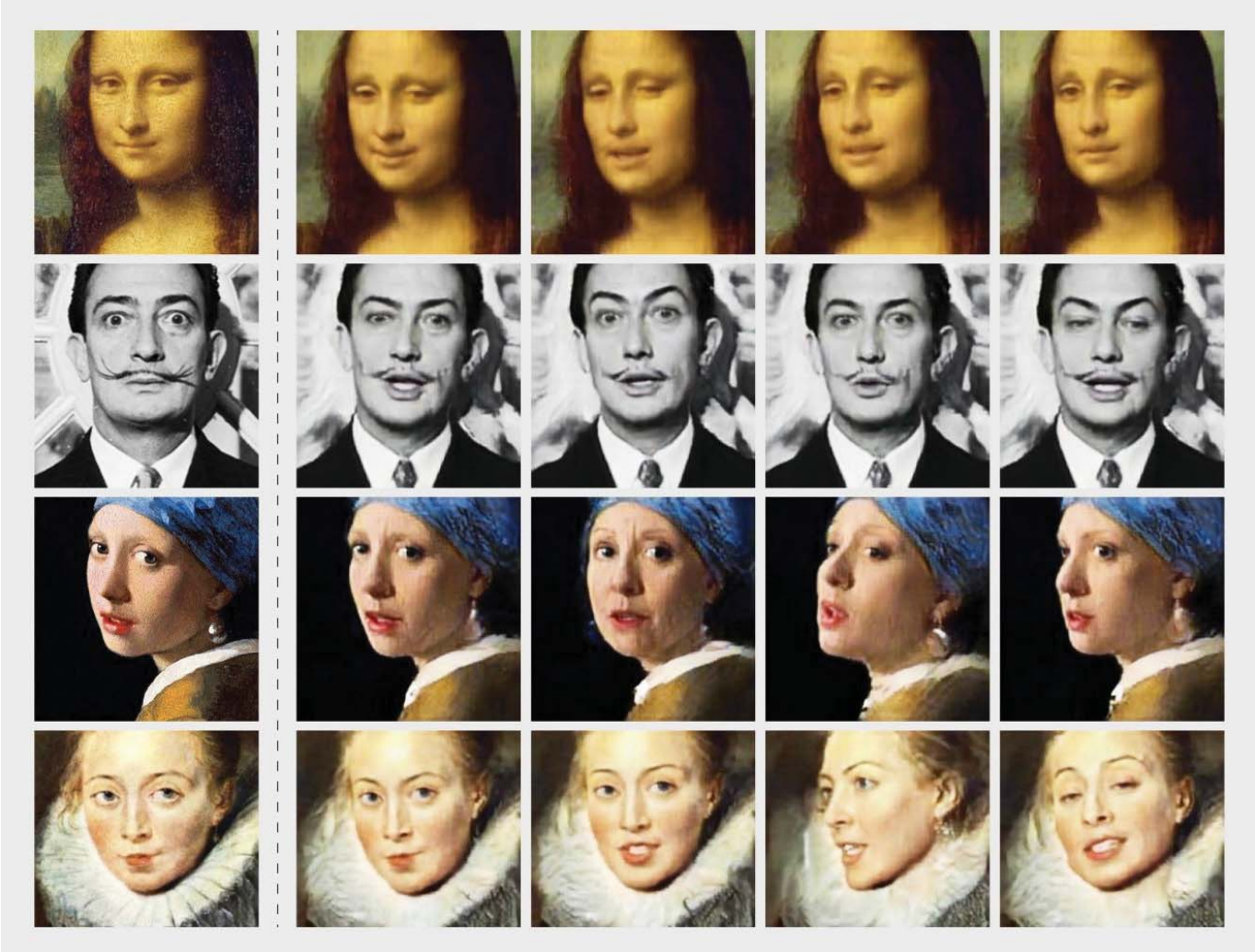
- detecting deepfakes is becoming harder as the technology becomes more advanced;
- determining the creator of a deepfake can be difficult;
- once the doubt has been raised, it is difficult to correct; and
- the speed with which these deepfakes can be disseminated far outstrips the speed with which their origin can be tracked.

While many have cried out for regulatory intervention, striking the right balance between competing rights and ensuring any legislation is agile enough to make an impact are challenges which need to be overcome in order to adequately address the concerns raised about this rapidly developing technology.

For better or worse, deepfakes currently occupy a place at a complex legal crossroads.

Intellectual property and deepfakes

One of the most obvious areas of law relevant to deepfakes is that of copyright. However, there are a number of complex issues which are likely to arise in this context. First, it must be noted that only copyright owners are able to sue for infringement and the owner of the copyright is not always the person who is the subject of the copyright work. This



Examples of deepfakes used to change the expressions or poses of famous artworks

means that while copyright may be one of the most obvious avenues for legal redress for those who have been the subject of a non-consensual deepfake, practically speaking, it may not be one that an aggrieved party is able to pursue personally. Assistance would be needed from the person or entity who owns the copyright, which in many cases where footage has been taken from movies or other large-scale productions would require major production companies agreeing to pursue copyright claims on behalf of those affected.

Another copyright issue that arises in the context of deepfakes relates to the ownership of the copyright in the deepfake itself. With the need for extensive input from AI and machine learning, attributing copyright ownership for media produced using a deepfake is beset by similar issues which arise in the context of AI generated artworks. At present, copyright protection is only afforded to works that have a human author.

Did Bruce Willis actually sell the rights to his face?

Returning to the example we started with, subsequent clarification from Willis's team served to highlight the complexities and uncertainties that surround this controversial application of AI technology. One question that remains is what precisely has Willis agreed to and what potential other uses could his digital twin be put to? Both Deepcake and Willis's team have confirmed that the actor has not sold the rights to his face to the deepfake producer. Rather, Deepcake claims to have been given Willis's consent to the creation of his deepfake twin and the various materials from which it drew the information base used to create the version of Bruce Willis which appears in the Megafon ad; however it appears that there is no agreement between the company and Willis himself.

This raises the question: who owns the rights to the digital twin.

Both the actor's media representatives and the deepfake generators have subsequently confirmed that all the rights in both his image and the digital twin belong exclusively to Bruce Willis, with the company neither retaining nor gaining any rights in its deepfake. It has further been revealed that Deepcake was not aware of the contractual arrangements between Megafon and Willis.

Fake it 'til you make it: tips and tricks for businesses

Deepfakes offer exciting new opportunities for both business and individuals to generate content that can be deployed in a myriad of ways, but as with many other rapidly developing technologies, the legal and regulatory frameworks are lagging far behind. With no specific legislation currently in force or contemplated by the Australian Parliament, businesses and individuals are left to navigate their way through the complex patchwork of existing laws, some of which offer more protection than others. As the recent US example has shown, however, implementing specific legislation to address the legal implications of deepfakes is not a simple or straightforward exercise, further complicated by the fact that the major players in this industry are based outside of Australia and so would likely escape the reach of any Australian legislation.

While we wait for the law to catch up, stories such as Bruce Willis' are only likely to increase, as more and more turn to this exciting technology to explore what can be achieved. However, deepfakes are not, however, only a problem for high-profile individuals, and businesses and individuals should also become familiar with some of the easy ways to spot a deepfake. While the technology behind deepfakes has developed rapidly and the outputs are now quite convincing, there are still a few key tell-tale signs that you are looking at a fake, rather than an original piece of content.

When it comes to faces, some of the key giveaways are:

- errors around the ears or the edges of the face;
- foreheads which appear unnaturally flat or which have oddly positioned wrinkles;
- facial expressions which do not match the tone of what is being said; and
- mouth movements which do not match what is being said or a mismatch between the mouth movements and the words being spoken.

When it comes to vocal deepfakes (either with images or just as a sound recording), some key giveaways are:

- odd pronunciation of words or phrases;
- use of words or phrases which are not characteristic for the speaker;
- comments which seem out of character for the person; and
- mismatches in the flow of the sound, caused by piecing together sounds which do not ordinarily occur together in normal speech patterns.

Finally, when it comes to spotting a deepfaked document, some key elements to check are:

- the format of the document (if it is an image, ask for the native file, not a PDF);
- check any payment details are correct by other means (including over the phone);
- where available, check the metadata for the file; and
- make sure all staff know how and when to escalate documents which they suspect are not legitimate.