Diving into the Deep End: Regulating Deepfakes Online

Ted Talas and Maggie Kearney, Ashurst, take us through the legal framework regulating deepfakes.

The term deepfake refers to a piece of video content that has been digitally manipulated using artificial intelligence. This technology can be used to seamlessly combine content from different sources, for example by superimposing a person's face onto a figure in a video as if the face were a mask. A deepfake can turn a person into a virtual ventriloquist's dummy, being made to appear as though they have said or done things that they have never said or done. The term deepfake, a portmanteau of "deep learning" and "fake", comes from the username of the Reddit poster that first began posting these videos online.

The potential consequences of deepfakes are significant: political leaders can be placed in compromising (and potentially election-losing) positions or even be shown to declare war on another country; celebrities can be used to endorse products or appear nude without their consent; consumers could be subject to complex phishing scams. At the same time, deepfakes can be used for less nefarious purposes, including parody, satire and entertainment (or however you would characterise inserting Nicolas Cage into every movie ever made[1]).

While the digital manipulation of visual content (whether using Photoshop or Instagram filters) is not a new phenomenon, the seamless manipulation of this content using artificial intelligence techniques is likely to present a more fundamental challenge to how we distinguish between what is fake and what is real.

In this article, we look briefly at how three areas of Australian law – copyright, defamation and the Australian Consumer Law (**ACL**) – could be used as tools to regulate the use and spread of deepfakes online. We have not attempted to provide an exhaustive list of the legal frameworks that may apply to deepfakes and acknowledge that many more could apply, including privacy laws and laws dealing with the non-consensual sharing of intimate images.

How do deepfakes work?

At this point in time, deepfakes are generally created using a machine learning framework called a generative adversarial network (**GAN**).[2] A GAN relies on two algorithms set to compete against each other. The first algorithm generates artificial samples of whatever you are trying to fake. These samples may be based off input data (eg, a collection of images of a particular person's face) or even random noise. The second algorithm compares these against a training data set to predict whether the sample is fake (ie, has been created by the generating algorithm) or real (ie, from the training data set). This process is then repeated (potentially many millions of times) with the predictions being fed back to the generating algorithm after each repetition to teach it how to make more and more realistic fakes which can be applied frame by frame to a video.[3]

While originally confined to academic work and the darker corners of the internet, since 2018, deepfake tools have been generally available online. This allows anyone with a set of photos or a video and sufficient computing power to create a deepfake (processing the repetitions required to make a convincing fake requires a relatively fast computer).

While deepfakes rely on artificial intelligence, it is important to keep in mind that it is possible to effectively manipulate videos without the use of this technology. For example, in May 2019, a video of Nancy Pelosi was released online which had been slowed so as to make Speaker Pelosi appear drunk.[4] These kinds of videos, which sometimes referred to as "cheapfakes", may be as problematic as deepfakes (for example, the Pelosi video was re-tweeted by President Trump).

Copyright

As outlined above, deepfakes ordinarily involve or incorporate existing video or audio content. Assuming this content is original, it is likely to be protected under the *Copyright Act 1968* (Cth) and copyright in the footage or recording will generally be owned by the person that made the film or recording (or their employer).

1 The A.V. Club, *Deep learning technology is now being used to put Nic Cage in every movie* (29 January 2018) <https://www.avclub.com/deep-learning-technology-is-now-being-used-to-put-nic-c-1822514573?rev=1517249018178&utm_content=Main&utm_campaign=SF&utm_source=Twitter&utm_medium=SocialMarketing>.

2 For further background, see Skymind, *A Beginner's Guide to Generative Adversarial Networks (GANs)* <https://skymind.ai/wiki/generative-adversarial-network-gan>.

3 For a useful illustration, see Australian Broadcasting Corporation, *How hard is it to make a believable deepfake?* (28 September 2018) <https://www.abc.net.au/news/2018-09-28/fake-news-how-hard-is-it-to-make-a-deepfake-video/10313906>.

4 Australian Broadcasting Corporation, *Nancy Pelosi speech manipulated to make her appear 'drunk' does not violate Facebook rules* (24 May 2019) <https://www.abc.net.au/news/2019-05-24/nancy-pelosi-speech-altered-video-slurring-words/11148030>.

If a deepfake reproduces a substantial part of the underlying film or recording, the owner of the film or recording would conceivably have a cause of action in copyright infringement against the creator of the deepfake and any person that subsequently reproduces or communicates it (subject at least to any fair dealing exception). A copyright claim may also arise in relation to the images or recordings used as the input to create the deepfake. As a part of this, the copyright owner could approach a Court to obtain an injunction to require the removal of the deepfake together with damages or an account of profits. The copyright owner could also seek to have the content removed under the takedown systems maintained by online platforms like Facebook.

The key limitation with using copyright to regulate deepfakes is that a copyright claim does nothing to vindicate – or even recognise – the damage caused by a deepfake to the person targeted by it, ie, the individual whose face and identity are used without their consent. Indeed, the person who is the target of the deepfake (and therefore likely to suffer the most harm as a result of its dissemination) is unlikely even to have standing to bring a claim for infringement or seek an injunction on copyright grounds. This is because, in the majority of cases, the target of a deepfake will not be the owner of the copyright in the underlying film. For example, if deepfake techniques are used to transplant a person's face into a pornographic video, only the maker of the video may be able to bring an infringement claim and not the person whose face was digitally inserted into the film.

Of course, where the interests of the deepfake target and the copyright owner align, this may not be an obstacle to removal of the deepfake. In June 2019, Condé Nast, the publisher of Vogue, successfully used YouTube's takedown request process to have a deepfake of Kim Kardashian West removed from the platform. The deepfake, which was created by a group of artists to lampoon influencer culture and would probably constitute fair dealing under Australia law, was based on footage from an interview Ms Kardashian West had done for Vogue in April 2019 (in which Condé Nast would have owned the copyright).[5]

However, these cases are likely to be rare, particularly for ordinary people without Ms Kardashian West's social media following. For most people, it may be impossible to identify a person whose copyright may be infringed by a deepfake. Even if such a person did exist and was identifiable, there is no guarantee they would be willing to assist the target of a deepfake by enforcing their rights as a copyright owner.

Another limitation with relying on copyright to deal with deepfakes, or indeed any of the legal frameworks discussed in this article, is that in many circumstances the creator of a deepfake may either be anonymous or located outside the jurisdiction of an Australian court. While the target of a deepfake may still be able to bring an action against an intermediary, such as an ISP or online platform, to have the content removed, these remedies may only be available in certain circumstances. This kind of intermediary liability may also have other unintended consequences, particularly given that intermediaries may face a commercial incentive to block content following a complaint rather than assessing the legitimacy of the complaint (eg, in circumstances where the alleged "deepfake" is actually real footage) or considering the applicability of any fair dealing exception.

As a result, while the law of copyright may be a useful tool to combat deepfakes in certain circumstances (particularly if the complainant is able to rely on online platforms' existing copyright takedown systems) the limitations discussed above mean that copyright law is unlikely to be a sufficient tool to address the proliferation of deepfakes online.

Defamation

Unlike copyright, the tort of defamation is specifically concerned with vindicating a person's reputation. It is not difficult to imagine a deepfake in which an identifiable individual is put into a compromising position or a scenario which could damage their reputation. The possibilities are literally endless. In these circumstances, and assuming the resulting video is made available to a third party, the target of a deepfake may be able to bring a defamation claim against any person involved in the publication of the deepfake to redress the damage to the target's reputation.

While the law of defamation has historically focused on words, whether spoken (slander) or written (libel), and is therefore well suited to address defamatory statements made in a deepfake, the law has recognised that images too can be defamatory. Famously, in *Ettingshausen v Australian Consolidated Press Ltd* (1991) 23 NSWLR 443, the New South Wales Supreme Court held that a photograph in which Mr Ettingshausen's genitals were apparently exposed was capable of subjecting the football player to ridicule and, therefore, of being defamatory. As a result, defamation law could also be used to provide a remedy in relation to the visual aspect of deepfakes.

Defamation law has also been applied to images that have been digitally altered. For example, *Charleston v News Group Newspapers Ltd* [1995] 2 AC 65 concerned an article published in *The News of the World*. The article, under the headline "Strewth! What's Harold

5 Vice Motherboard, *The Kim Kardashian Deepfake Shows Copyright Claims Are Not the Answer* (20 June 2019) <https://www.vice.com/en_us/article/j5wngd/kim-kardashian-deepfake-mark-zuckerberg-facebook-youtube>.

up to with our Madge", included a large photograph of a man and a woman nearly naked and apparently engaging in sexual activity with the faces of actors from the television soap *Neighbours* superimposed on each body. The actors sued for defamation, including on the basis that they had been made out as willing participants in the creation of the photograph.

In this case, the House of Lords dismissed the actors' claim, holding that the publication was incapable of conveying any of the defamatory meanings pleaded. This was because any defamatory sting in the photograph was effectively neutralised by the accompanying text in the article which clarified that the photograph had been produced by the makers of a pornographic computer game without the knowledge or consent of the actors.

This is not to suggest that the mere fact that a deepfake includes a disclaimer that it is the product of digital manipulation, or even if this is apparent from the poor quality of the video, will frustrate a claim for defamation. It will all depend on the imputations pleaded. For example, Senator Sarah Hanson-Young successfully brought defamation proceedings against Zoo magazine in relation to an article featuring a plainly photo-shopped image featuring Senator Hanson-Young's face on the head of a bikini model (the imputations included that the article suggested that Senator Hanson-Young was not a serious politician).

Given this area of the law focuses on vindicating a person's reputation regardless of how they are defamed, the established principles of defamation appear to be well suited to addressing the reputational harm caused by deepfakes, particularly given that Courts in Australia have a record of applying the established principles of defamation law in new online contexts.

However, although defamation may provide a mechanism for the target of a deepfake to obtain compensation, in certain circumstances, it may be a less effective mechanism to force the removal of deepfakes. This is because Australian Courts are often reluctant to grant injunctions in defamation proceedings due to free speech concerns, particularly on an interlocutory basis. However, this reluctance may not apply in circumstances where there is no public interest in the deepfake remaining available online (eg, in the context of revenge porn).

Australian Consumer Law

Unlike other jurisdictions, the common law of Australia does not recognise an independent cause of action to protect how a person's identity, including their name and likeness, is used. In jurisdictions where these publicity rights are recognised, claims based on such rights are likely to be a useful tool to be deployed against deepfakes, at least where the deepfake is used in a commercial context.

Nevertheless, a plaintiff in Australia may be able to rely on the ACL in a similar way. By way of illustration, consider a deepfake in which a famous tennis player was made to endorse, without their knowledge, a tennis racquet made by a manufacturer other than their sponsor. Such a video would potentially contravene the provisions of the ACL, for example, sections 29(1)(g), which prohibits a person, in connection with supplying or promoting goods, making a false or misleading representation that the goods have a certain affiliation, approval or sponsorship or the general prohibition on misleading and deceptive conduct in section 18.

The ACL provides for a wide range of remedies for any contravention, including injunctions and damages. Additional orders, including pecuniary penalties, are also available in enforcement actions brought by the ACCC. In addition to the ACL, our hypothetical tennis player may also be able to rely on the tort of passing off to obtain an injunction, damages or an account of profits in relation to the deepfake.

While the ACL appears to be an ideal tool to combat deepfakes (after all, deepfakes are, by definition, misleading and deceptive), in reality, its application is limited. This is because the ACL generally only applies to commercial activity. For example, the prohibition on section 18 only applies to conduct "in trade or commerce". The ACL is therefore only likely to capture deepfakes used to promote a product or service, criticise a business so as to influence consumer behaviour or where the deepfake itself is being directly monetised (eg, through the sale of online advertisements). These laws are unlikely to apply to deepfakes used in other, and potentially more insidious, contexts, such as revenge porn or political disinformation (although other more-targeted laws may apply in those scenarios).

Conclusion

While existing legal frameworks may be appropriate to regulate deepfakes in certain circumstances, these frameworks are unlikely to be sufficient to address the fundamental challenge that deepfakes pose to society.

There is no doubt a role to play for new laws dealing with the spread of deepfakes and other disinformation online.[6] However, our view is that future legislative reform will only ever form part of an effective solution. What is required is the continuing development of effective tools to detect, identify and alert internet users of deepfakes. Poetically, many of these tools rely on the artificial intelligence that makes deepfakes possible in the first place.

6 For example, the ACCC has recently proposed the creation of an industry code to govern the handling of complaints in relation to the spread of disinformation on digital platforms (and which would capture deepfakes). See Australian Competition and Consumer Commission, *Digital Platforms Inquiry* (Final Report, June 2019) 370.